AD_____

Award Number: DAMD17-00-1-0108

TITLE: An LOH Study of Chromosome 8 in Multiplex Prostate Cancer Sibships

PRINCIPAL INVESTIGATOR: Brian K. Suarez, M.D., Ph.D.

CONTRACTING ORGANIZATION: Washington University
Saint Louis, Missouri 63110

REPORT DATE: September 2001

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 2001	Annual (1 Mar 00 - 31 Aug 01)

4. TITLE AND SUBTITLE
An LOH Study of Chromosome 8 in Multiplex Prostate Cancer Sibships

5. FUNDING NUMBERS
DAMD17-00-1-0108

6. AUTHOR(S)
Brian K. Suarez, M.D., Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Washington University
Saint Louis, Missouri 63110

E-Mail: bks@themfs.wustl.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

Purpose: This study is designed to identify one or more prostate cancer (CaP) tumor suppressor genes (TSG) on chromosome 8 by assessing the distribution of loss of heterozygosity in brothers from multiplex sibships. A strong linkage signal has been demonstrated in these families. The differential pattern of linkage disequilibrium between affected men and controls will be used to further narrow the region and the potential for epistatic interaction between a TSG and two non-chromosome 8 putative CaP susceptibility genes (HPC2/ELAC2 and the androgen receptor) will be assessed with methods currently under development.
Scope: The region of interest is the entire short arm of chromosome 8 and a region from approximately D8S2324 to D8S592 on 8q.
Major Findings: Although the genotyping is not yet complete, we find levels of LOH in our multiplex sample comparable to levels reported among unrelated men. The region of greatest LOH stretches from D8S277 to D8S1825 and corresponds to the peak linkage signal.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Gene Discovery, Tumor Suppressor Genes, Loss of Heterozygosity, Linkage, Multiplex Prostate Cancer Sibships	91
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

INTRODUCTION: The purpose of this study is to conduct a search for a tumor suppressor gene (TSG) or genes on chromosome 8 that when inactivated increase a man's susceptibility to develop prostate cancer (CaP). Our approach is to identify regions on chromosome 8 that display a similar pattern of loss of heterozygosity (LOH) in brothers with CaP. Additionally, since it is possible that alleles of a TSG will be in disequilibrium with alleles at chromosome 8 microsatellite markers (used to detect linkage) and/or epistatically interact with alleles at other (unlinked) susceptibility loci, we have also genotyped 2 non-chromosome 8 genes alleged to be etiologically involved with CaP development and have begun developing methods that could aid in the identification of TSGs under these conditions.

BODY: DNA from a total of 157 men from multiplex prostate cancer (CaP) families has been genotyped to date. The number of completed genotypes is 776. In order to be "complete" the marker must successfully be genotyped in both tumor and constitutional DNA. Table 1 reports the microsatellites that have been genotyped, their level of heterozygosity in our data and the level of LOH observed to date.

Table 1

Marker	Percent Heterozygosity	Percent Informative with LOH
D8S1781	87.5	0.0
D8S262	58.1	38.9
D8S277	71.9	52.2
D8S351	79.1	45.3
D8S1825	66.2	46.8
D8S1130	84.1	33.3
D8S1106	74.2	33.7
D8S1731	85.6	36.8
D8S261	68.6	45.8
D8S560	71.4	50.0
D8S1144	68.2	26.7
D8S175	62.9	27.3
D8S1822	88.2	36.7
D8S521	52.8	21.1

The sample size at marker D8S1781 is small so it is likely that as new samples are genotyped, the percent LOH will increase. Table 2 reports the distribution of the 6 possible patterns for an affected pair of brothers observed in our affected sib pairs to date.

Table 2

Pattern	Percent
Homozygous/Homozygous	13.6
Homozygous/No LOH	18.3
Homozygous/LOH	5.9
No LOH/No LOH	31.4
No LOH/LOH	24.9
LOH/LOH	5.9

Our original hypothesis remains the same; namely that the pattern of LOH (with respect to which allele or haplotype is retained) will be correlated within sibships as a function of the sibship's identity-by-descent (IBD) configuration. This hypothesis logically follows from (and tests the assumption that) the excess allele sharing revealed by our multipoint linkage analysis is due to the presence of a shared "mutated" tumor suppressor gene in a proportion of these sibships. When genotyping of constitutional and tumor DNA of all subjects in this study is completed, we will obtain multipoint estimates of the probability that an affected sib pair (ASP) shares zero, one or both alleles IBD for each microsatellite. The greater the density of microsatellites, the greater the accuracy of the IBD inferences. For reasons given below, the sample affected brothers genotyped to date is too small to carry out these analyses at this time.

A no-cost one-year extension of this study was sought in July and approved in August, 2001. The reason the extension was requested is because we experienced unanticipated delays in obtaining archived tumor tissue. There are many steps in this process. For subjects who had their treatment (usually a radical prostatectomy) here at Barnes Hospital, a request for the paraffin block must be made by the participating pathologist. Retrieval can be a lengthy process. If treatment was performed elsewhere, a request for transfer of a paraffin block must be made of the source hospital. Our requests were not always treated with a sense of urgency. Once the block is received it is sectioned, mounted and stained. The sections are examined by a pathologist and those that contain sufficient tumor were sent to colleagues in Pittsburgh or Denver who prepared the laser capture microdissected (LCMD) caps. (Not all blocks contain tumor and when this occurs, additional blocks need to be requested, causing a further delay). Our experience is that the amount and quality of tumor DNA varies widely from cap-to-cap. Under the best conditions, our genotyping laboratory was able to perform about 8 PCR reactions per cap although the modal number is fewer.

As soon as it became clear that the amount of tumor DNA was limiting the progress of this project we decided to pursue another strategy. To this end we shipped caps to a laboratory in Cincinnati with expertise in performing whole genome amplification (WGA) of minute archival DNA samples to determine if WGA could be used on our LCMD caps to increase the yield of tumor DNA. These experiments succeeded and an agreement has been formalized for this lab to perform WGA on our caps and to genotype the microsatellites shown in Table 1. There is an additional important advantage to this new collaboration. A proportion of our tumor samples exist only as needle biopsies. We do not know if these biopsy specimens contain a sufficient number of tumor cells to allow WGA, but if they do, we will be able to substantially increase our projected sample size by making use of this technology. The whole genome amplification will be done following the

PEP (Primer Extension preamplification) protocol using the High Fidelity Expand System. We have secured support from a private research foundation for this work.

Because we had little control over the accrual process, we were unable to genotype members of the same sibship as a unit. Thus, for seventeen of our affected trios, only 1 of the 3 brothers has been genotyped, and for another eight trios, only 2 of the 3 brothers have been genotyped. Accordingly, an additional 25 completed trio families will be added to our sample once tumor DNA is obtained from the outstanding 42 subjects. The one-year extension of this project should allow us to fill in these gaps.

Non-Chromosome 8 Genotypes: We elected to genotype 3 polymorphisms in two non-chromosome 8 genes that have been alleged to affect susceptibility to CaP. Since either of these non-chromosome 8 genes could interact with one or more chromosome 8 TSGs, it seemed prudent to be in a position to control (at least statistically) for possible interactions. Accordingly, we genotyped a non-synonymous coding SNP in the newly discovered CaP susceptibility gene, HPC2/ELAC2, on chromosome 17 (Tavtigian et al., 2001). We published the results of this work earlier this year (Suarez et al., 2001) and although we did not find a main effect for this polymorphism, we now have the data to assess its possible interaction with chromosome 8 markers once the on-going genotyping is completed. Additionally, we genotyped two trinucleotide repeat polymorphisms in exon 1 of the X-linked androgen receptor gene (AR). Variation in these repeats is widely believed to affect risk for the development of CaP. A manuscript (Suarez et al., submitted) describing the results of this work is currently under review. As with the HPC2/ELAC2 genotypes, the AR genotypes will be available for epistatic analysis (see below) once the chromosome 8 genotyping is complete.

Theoretical Work: Linkage disequilibrium in a population refers to the nonrandom association of alleles at 2 (or more) loci on the same chromosome. We hypothesize that once the chromosome 8 genotyping is completed, analysis of the LOH patterns within sibships will delineate one or more subregions as candidates for harboring a TSG. It is possible, however, that these regions will be too large to undertake positional cloning. One possible method to delimit a region would be to show greater linkage disequilibrium in CaP cases than in controls. To this end, we have collected a large series of race-matched controls at no cost to the present study. These controls will be used to help narrow the region of interest using a new statistic developed for this purpose. A paper describing the statistic and its application to a complex phenotype in simulated data has just been published (Culverhouse et al., 2001).

It is becoming clear that many, perhaps most, complex phenotypes such as prostate cancer result from the interaction of two or more loci. Since a large number of linkage "signals" for CaP have been reported but no replicated susceptibility gene has been identified, it is conceivable that the genes involved either have no "main effect" or a main effect too small to be detected by standard association methods. To explore this possibility, we have developed an approach that maps the limits of purely epistatic models, and are working on new statistical methods for the detection of component loci. We will test epistasis between the chromosome 8 markers and the two non-chromosome 8 candidate genes (HPC2/ELAC2 and the AR) with these new methods. A manuscript (Culverhouse et al., submitted) describing this work is under review.

KEY RESEARCH ACCOMPLISHMENT:

- Fourteen chromosome 8 microsatellites were genotyped in a total of 157 men from multiplex prostate cancer families.
- The percent of informative markers with LOH ranges from 0% for the telomeric microsatellite D8S1781 to 52.2% for D8S277.
- 31.4% of informative markers in pairs of brothers show no LOH, while 24.9% reveal LOH in one of the brothers but not the other. For 5.9% of the markers, both brothers show LOH.
- We have developed a new statistic that exploits differences in patterns of linkage disequilibrium between cases and controls that should help delimit the position of a TSG on chromosome 8.
- We have developed a method that maps the limits of purely epistatic models and are developing techniques capable of identifying participating susceptibility loci.

REPORTABLE OUTCOMES:

The following manuscripts have been published or are under review. Copies are included in the appendix.

1. Suarez BK, Gerhard DS, Lin J, Haberer B, Nguyen L, Kesterson NK, and Catalona WJ: Polymorphisms in the prostate cancer susceptibility gene HPC2/ELAC2 in multiplex families and healthy controls. Cancer Res. 61:4982-4984 (2001).
2. Culverhouse R, Lin J, Liu K-Y, and Suarez BK: Linkage disequilibrium mapping in population isolates. Genet. Epid. 21:S429-434 (2001).
3. Suarez BK, Lin J, Catalona WJ, Haberer B, and Gerhard DS: CAG and GGC trinucleotide repeats in the androgen receptor gene and prostate cancer: The long and the short of it. Submitted.
4. Culverhouse R, Suarez BK, Lin J, and Reich T: A perspective on epistasis: Limits of models displaying no main effect. Submitted.

CONCLUSIONS:

Prostate cancer is the most common malignancy and second leading cause of cancer-related death among American men. It is estimated that this year, 198,000 men will be newly diagnosed and about 31,300 will die of the disease (Greenlee et al., 2001). There is now a persuasive body of evidence implicating one or more tumor suppressor genes on chromosome 8. This evidence derives from many LOH studies of unrelated men with prostate cancer. Because these studies are difficult, most findings have been based on relatively small sample sizes. We have adopted a different strategy in this study, namely to investigate the pattern of LOH in related subjects from multiplex sibships. We believe this will be the first LOH study of related men with prostate cancer.

In the event that the regions are still too large to contemplate positional cloning, we have developed an approach that exploits differential linkage disequilibrium in cases versus controls. Additionally, we have characterized the genetic variation at two non-chromosome 8 putative CaP susceptibility genes in our sample (HPC2/ELAC2 and the AR) with the expectation of undertaking an analysis of epistasis between these genes and the yet-to-be identified chromosome 8 TSG.

REFERENCES:

1. Culverhouse R, Lin J, Liu K-Y, and Suarez BK: Linkage disequilibrium mapping in population isolates. Genet. Epid. 21:S429-434 (2001).
2. Culverhouse R, Suarez BK, Lin J, and Reich T: A perspective on epistasis: Limits of models displaying no main effect. Submitted.

3. Greenlee RT, Hill-Harmon MB, Murray T, and Thun M: Cancer statistics, 2001. Cancer J. Clin. 51:15-36 (2001).
4. Suarez BK, Gerhard DS, Lin J, Haberer B, Nguyen L, Kesterson NK, and Catalona WJ: Polymorphisms in the prostate cancer susceptibility gene HPC2/ELAC2 in multiplex families and healthy controls. Cancer Res. 61:4982-4984 (2001).
5. Suarez BK, Lin J, Catalona WJ, Haberer B, and Gerhard DS: CAG and GGC trinucleotide repeats in the androgen receptor gene and prostate cancer: The long and the short of it. Submitted.
6. Tavtigian SV, Simard J, Teng DHF, Abtin V, Baumgard M, Beck A, Camp NJ,Carillo AR, Chen Y, and Dayananth P: A candidate prostate cancer susceptibility gene at chromosome 17p. Nat. Genet. 27:172-180 (2001).

[CANCER RESEARCH 61, 4982–4984, July 1, 2001]

Appendix
DAMD17-00-1-0108

Advances in Brief

Polymorphisms in the Prostate Cancer Susceptibility Gene *HPC2/ELAC2* in Multiplex Families and Healthy Controls[1]

Brian K. Suarez,[2] Daniela S. Gerhard, Jennifer Lin, Beth Haberer, Loan Nguyen, Niki K. Kesterson, and William J. Catalona

Departments of Psychiatry [B. K. S., D. S. G., J. L.], and Genetics [B. K. S., D. S. G., L. N., N. K. K.], and Division of Urologic Surgery [B. H., W. J. C.], Washington University, School of Medicine, St. Louis, Missouri 63110

Abstract

Two polymorphisms in the newly cloned prostate cancer susceptibility gene, *HPC2/ELAC2*, are suspected to be associated with an increased risk of developing the disease. These missense variants result in a serine (S) to leucine (L) substitution at amino acid residue 217 and an alanine (A) to threonine (T) substitution at residue 541. We genotyped these polymorphisms in 257 multiplex prostate cancer sibships and in 355 race-matched healthy unrelated controls. A significant increase in the frequency of the *T* allele is seen in the prostate cancer subjects compared with controls. There is, however, little evidence for excess clustering of the *T* allele within the multiplex families known to be segregating this allele, and there is no evidence for linkage of prostate cancer to the *HPC2/ELAC2* region of chromosome 17p11.2 in these families. The *T* allele shows no association with either Gleason score or age-of-onset in segregating families.

Introduction

Recently, Tavtigian *et al.* (1) announced the positional cloning of the first CaP[3] susceptibility gene (*HPC2/ELAC2*) to be identified as the result of a whole genome scan of high risk CaP pedigrees. Two-point linkage analysis performed on a subset of the families yielded a heterogeneity lod score of 4.43 at *D17S1289*. Analysis of recombinants in two Utah kindreds allowed narrowing of the interval to about 1 Mb. Subsequent analysis of the genomic sequence across this region revealed two independent multiexon genes, *04CG09* and *HPC2/ELAC2*. Mutation screening of genomic DNAs from an early age-of-onset multiplex CaP family revealed a frameshift mutation in *HPC2/ELAC2* that throws translation out of frame after amino acid 547 and that cosegregated with the CaP phenotype (1). In addition, two missense variants, a serine (S) to leucine (L) substitution at amino acid 217 and an alanine (A) to threonine (T) substitution at amino acid 541, were found to be associated with the diagnosis of CaP. Rebbeck *et al.* (2) confirmed this association in a sample of 359 incident CaP cases and 266 matched controls. The association was attributable to a significant increase of the *T* allele at amino acid residue 541 in the CaP cases compared with controls.

Materials and Methods

We genotyped these polymorphisms in 257 multiplex CaP sibships and 355 unrelated controls. All of the multiplex sibships were ascertained from patients seen at Washington University School of Medicine by staff urologists, or were referred by other area urologists, or were participating in CaP support groups, or responded to our published solicitations. Two hundred and thirteen of these families were included in our initial genome scan (3), 22 were added to our follow-up linkage study of chromosomes 1 and 16 (4), and 22 are new to this study.

The control subjects were ascertained from a large pool of men who have been followed for many years as part of a long-term CaP study in which men are screened at 6-to-12 month intervals with PSA blood tests and DRE of the prostate (5). The size of this pool allowed us the luxury of identifying especially healthy men. To be enrolled in the control series, the subjects were required to meet the following four criteria: (*a*) be at least 65 years old; (*b*) never have registered a PSA level in excess of 2.5 ng/ml; (*c*) never have had a DRE suspicious of CaP; and (*d*) have no known family history of CaP. This last criterion was assessed by inquiring about the subject's brothers, father, grandfathers, and maternal and paternal uncles. As a consequence of the first criterion, the control subjects were significantly older than the case subjects (71.7 *versus* 65.5 years; $P < 0.0001$). All of the subjects in this study were of European ancestry. The protocol of this study was approved by the Human Studies Committee of Washington University, and informed written consent was obtained from all of the participants.

The S217L variation was detected as follows: 37 ng of genomic DNA was amplified in 8 μl of total volume using standard 1.5 mM $MgCl_2$ polymerase buffer and 1.25 μM each of 217*HPC2/ELAC2*-5′ (CAGCTCACCTTGTGCAGTGT) and 217*HPC2/ELAC2*-3′ (GCCCAGGAAGAAGGATCTGT) primers, and 0.1 unit of *Taq* polymerase. The DNA was first denatured at 94°C for 2 min and then amplified in 35 cycles of 92°C for 30 s, 63°C for 1 min, and 72°C for 1 min. Each PCR product of 294 bp was digested with 2 units of *TaqI* (New England Biolabs, Inc.) at 65°C for 2.5–3 h. One-half of the DNA was electrophoresed at 130 V for 1.5 h on 3% agarose 3:1 (Amresco) to separate the fragments, 294 (L) *versus* 157 and 137 (S).

The A541T variation was detected by amplification of 37 ng of genomic DNA in 15 μl of final volume using the same conditions as for S 217 L, except that the primers were 541*HPC2/ELAC2*-5′ (CCTGTCCAAAGCAGACATCA) and 541*HPC2/ELAC2*-3′ (AGGAAAAGACGCAGCCAAAG), and the annealing temperature was 60°C. Each PCR product of 303 bp was digested with 1 unit of *Fnu 4HI* at 37°C for 3 h. The electrophoretic conditions were the same. All of the alleles had two visible constant bands (the 13-bp product could not be separated from the primers) of 79 and 49 bp. These bands were the positive control for the enzyme function. The 110-bp and 52-bp fragments (the latter comigrated with the 49-bp product) were diagnostic of the *A* allele, and the 162-bp fragment was diagnostic for the *T* allele.

All of the allele calls were independently verified by two of the authors, D. S. G. and either L. N. or N. K. K. Additionally, 15% of the DNA samples were regenotyped, and no discrepancies were observed.

Results and Discussion

Table 1 reports the joint genotype distribution for the two *HPC2/ELAC2* variants in 355 unrelated controls and 257 CaP cases. The genotypic distribution for the cases was obtained by sampling at random a single brother from each of 257 multiplex sibships and repeating the process 1000 times. The mean from these 1000 random samples is reported in Table 1. Similar to the findings of Rebbeck *et al.* (3), we observe complete linkage disequilibrium between the *T*

Received 2/26/01; accepted 5/16/01.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] Supported in part by awards from the Urological Research Foundation, the CaP CURE Foundation, USPHS Grant MH31302, and DAMD17-00-0108 from the United States Army.

[2] To whom requests for reprints should be addressed, at Department of Psychiatry, Washington University, School of Medicine, Campus Box 8134, 660 S. Euclid, St. Louis, MO 63110. Phone: (314) 362-9433; Fax: (314) 747-1017; E-mail: bks@themfs.wustl.edu.

[3] The abbreviations used are: CaP, prostate cancer; PSA, prostate-specific antigen; DRE, digital rectal examination; ASP, affected sib pair.

Table 1 *Joint genotype distribution of the S/L polymorphism at amino acid 217 and the A/T polymorphism at position 541 in a random sample of 257 unrelated prostate cancer cases and 355 healthy unrelated controls*

The joint genotype distribution for the cases was obtained by sampling at random one brother from each sibship. The reported number for each genotype (rounded to the nearest integer) is the mean based, on 1000 realizations of the sampling procedure. Genotypic proportions are given in parentheses.

	Residue 541					
	Cases			Controls		
Residue 217	*AA*	*AT*	*TT*	*AA*	*AT*	*TT*
SS	120 (0.467)			190 (0.535)		
SL	93 (0.362)	21 (0.082)		124 (0.349)	10 (0.028)	
LL	19 (0.074)	4 (0.016)		28 (0.079)	1 (0.003)	2 (0.006)
Allele frequency:	$\bar{L} = 0.3113$			$\bar{L} = 0.2761$		
	$\bar{T} = 0.0486$			$\bar{T} = 0.0211$		

allele and the *L* allele; accordingly, the maximum likelihood estimate of the *L-T* haplotype is the same as the estimate of the *T* allele frequency reported in Table 1. There is no significant difference in the frequency of the *L* and *S* alleles between cases and controls ($X^2 = 1.79$; $P = 0.18$). However, the frequency of the *T* allele is significantly greater in the CaP cases than the control sample ($X^2 = 7.13$; $P = 0.008$).

To determine whether the apparent increase of the *T* allele in our cases is attributable to excess clustering within families, we identified all of the sibships that contained at least one brother with at least one *T* allele. As can be inferred from Table 1, all of the *T*-bearing CaP cases in our sample were *A/T* heterozygotes. Forty families (33 ASPs and 7 affected trios) were identified. To test whether these 40 families provided evidence for excess clustering of the *T* allele, we computed the probability of observing particular genotypic arrays conditional on the observed sibship configurations. For instance, among the ASPs in our sample, only two configurations were observed: *AT,AT* pairs and *AT,AA* pairs. Letting p denote the frequency of the *A* allele and $q = 1 - p$, the frequency of the *T* allele, the proportion of ASPs expected to be *AT,AT* was:

$$\frac{1 + pq}{3p^2 + 4pq + q^2} \qquad \text{(A)}$$

Similarly, among our affected trios, only two configurations were observed: *AT,AT,AA* trios and *AT,AA,AA* trios. The proportion of affected trios expected to be *AT,AT,AA* was:

$$\frac{\frac{1}{2}p + \frac{1}{4}q}{p + \frac{3}{8}q} \qquad \text{(B)}$$

Table 2 reports these expectations for our sample of 33 ASPs and 7 affected trios for selected values of $q < 0.10$. The expected proportion of *AT,AT* ASPs or *AT,AT,AA* trios did not strongly depend on the estimated frequency of the *T* allele.

Among the 33 ASPs in our sample, 10 had configuration *AT,AT*. Three of the 7 trios had configuration *AT,AT,AA*. Because we observed slightly less clustering than expected by chance, these limited data did not support the hypothesis that the *T* allele is overrepresented in segregating CaP families.

The identification and cloning of the *HPC2/ELAC2* gene represents a significant milestone in the quest to understand the genetic architecture of hereditary CaP. The data presented here, however, appear paradoxical. Thus, whereas we were able to demonstrate an increased frequency of the *T* allele in CaP cases compared with controls, we did not find excess clustering in multiplex sibships as would be expected if this allele substantially increases risk. Our approach to detect excess clustering was similar in spirit to the family-based association test described in Parsian *et al.* (6) or the formal TDT test described by Spielman *et al.* (7), except that, because we did not genotype any parents, ascertainment was through *T*-bearing affected offspring. Thus, although our families did not offer a great deal of power to reject the hypothesis that the *T* allele accounts for a small increase in risk, neither did they provide even a hint of support.

It is of interest that none of the three published whole genome scans of CaP families (3, 8, 9) yield any compelling multipoint evidence of a susceptibility locus on chromosome 17p, although a weak two-point signal was reported by Gibbs *et al.* (9). We have reanalyzed our families for linkage (10, 11) by subdividing them into two groups depending on the presence/absence of the *T* allele. Chromosome 17p genotypes are available on 35 of our 40 *T*-positive sibships and on 200 of the 217 *T*-negative sibships. As can be seen in Fig. 1, the multipoint NPL *Z*-scores for neither subgroup give any evidence of increased allele sharing in the vicinity of *HPC2/ELAC2*.

The frequency of deleterious mutations in *HPC2/ELAC2* is unknown. The 1641 insG insertion reported by Tavtigian *et al.* (1), causes a frameshift in a region of the protein that displays a high degree of amino-acid-sequence conservation among multicellular eukaryotes. In their recent review, Ostrander and Stanford (12) suggested that germ-line changes in *HPC2/ELAC2* are unlikely to be a common cause of CaP because an examination of 45 additional unrelated CaP cases, all with an age-of-onset of 35–55 years, failed to reveal any with a frameshift mutation. Of course it is possible that variation in noncoding regions could alter the transcription of the *HPC2/ELAC2* gene. If such elements are identified and if they are in linkage disequilibrium with the *T* allele, then the association reported here and in the study of Rebbeck *et al.* (2) must be considered secondary and noncausal.

It is also possible that the significant case-control difference that we observed was attributable to having selected an unusually healthy control

Table 2 *Predicted number (of 33) of affected sib pairs with the AT,AT genotype configuration, and the predicted number (of 7) of affected trios to have the AT,AT,AA configuration*

The observed numbers are 10 and 3, respectively.

q^a	ASPs *AT,AT*	Trios *AT,AT,AA*
0.03	11.55	3.51
0.05	11.92	3.52
0.07	12.29	3.53
0.09	12.66	3.54

[a] q, the allele frequency of the *T* allele.

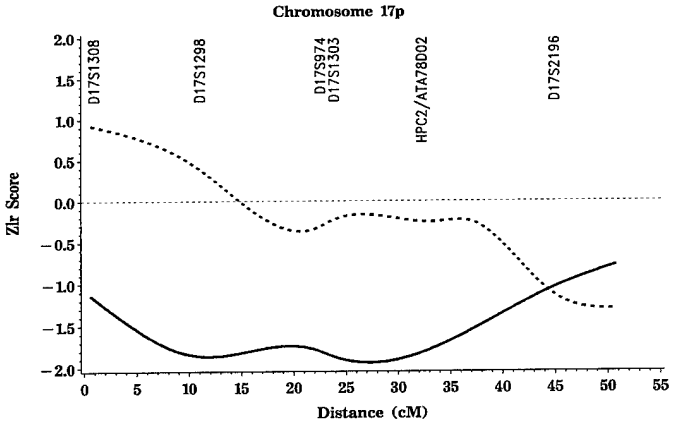

Fig. 1. Linkage analysis of chromosome 17p in 35 multiplex CaP sibships that contained at least one CaP brother with an *HPC2/ELAC2 T* allele (*solid line*) and in 200 multiplex sibships that contained only brothers who were homozygous for the *A* allele (*broken line*). GENEHUNTER-PLUS (10, 11) was used to compute the Zlr scores.

sample. Not only did all of our control men have serial PSA measurements that had never exceeded 2.5 ng/ml and consistently negative DRE findings, but none had a first- or second-degree relative with CaP. The frequency of the *T* allele in our controls (0.021) is lower than that reported by Rebbeck *et al.* (2) for their white controls (0.032), albeit the difference is nonsignificant ($X^2 = 1.41$; $P = 0.24$).

Finally, to determine whether the *T* allele affects either Gleason score or age-of-onset, we drew at random 40 unrelated *A/T* heterozygotes (one from each family) and compared them with an unrelated random sample of *A/A* homozygotes drawn from the remaining 217 families. These measurements were unavailable for one of the *A/T* subjects and four of the *A/A* subjects, thereby reducing the sample size slightly. No differences were seen for either mean Gleason score ($P = 0.94$) or mean age-of-onset ($P = 0.30$).

In summary, we found a significant increase in the frequency of the *T* allele in CaP cases drawn from multiplex sibships compared with exceptionally healthy race-matched controls. Analysis of these families, however, failed to reveal any excess clustering of the *T* allele as would be expected if this amino acid substitution substantially increased susceptibility. Furthermore, linkage analysis of families segregating the *T* allele does not suggest that this polymorphism is linked to CaP in this sample.

References

1. Tavtigian, S. V., Simard, J., Teng, D. H. F., Abtin, V., Baumgard, M., Beck, A., Camp, N. J., Carillo, A. R., Chen, Y., and Dayananth, P. A candidate prostate cancer susceptibility gene at chromosome 17p. Nat. Genet., *27:* 172–180, 2001.
2. Rebbeck, T. R., Walker, A. H., Zeigler-Johnson, C., Weisburg, S., Martin, A-M., Nathanson, K. L., Wein, A. J., and Malkowicz, S. B. Association of *HPC2* genotypes and prostate cancer. Am. J. Hum. Genet., *67:* 1014–1019, 2000.
3. Suarez, B. K., Lin, J., Burmester, J. K., Broman, K. W., Weber, J. L., Banerjee, T. K., Goddard, A. B., Witte, J. S., Elston, R. C., and Catalona, W. J. A genome screen of multiplex sibships with prostate cancer. Am. J. Hum. Genet., *66:* 933–944, 2000.
4. Suarez, B. K., Lin, J., Witte, J. S., Conti, D. V., Resnick, M. I., Klein, E. A., Burmester, J. K., Vaske, D. A., Banerjee, T. K., and Catalona, W. J. Replication linkage study for prostate cancer susceptibility genes. Prostate, *45:* 106–114, 2000.
5. Smith, D. S., Humphrey, P. A., and Catalona, W. J. The early detection of prostate cancer with prostate-specific antigen: the Washington University experience. Cancer (Phila.), *80:* 1852–1856, 1997.
6. Parsian, A., Todd, R. D., Devor, E. J., O'Malley, K. L., Suarez, B. K., Reich, T., and Cloninger, C. R. Alcoholism and alleles of the human D_2 dopamine receptor locus. Arch. Gen. Psychiatry, *48:* 655–663, 1991.
7. Spielman, R. S., McGinnis, R. E., and Ewens, W. J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (*IDDM*). Am. J. Hum. Genet., *52:* 506–516, 1993.
8. Smith, J. R., Freije, D., Carpten, J. D., Grönberg, H., Xu, J., Isaacs, S. D., Brownstein, M. J., Bova, G. S., Guo, H., Buinovszky, P., *et al.* Major susceptibility locus for prostate cancer on chromosome 1 suggested by a genome-wide search. Science (Wash. DC), *274:* 1371–1374, 1996.
9. Gibbs, M., Stanford, J. L., Jarvik, G. P., Janer, M., Badzioch, M., Peters, M. A., Goode, E. L., Kolb, S., Chakrabarti, L., Shook, M., Basom, R., Ostrander, E. A., and Hood, L. A genomic scan of families with prostate cancer identifies multiple regions of interest. Am. J. Hum. Genet., *67:* 100–109, 2000.
10. Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., and Lander, E. S. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am. J. Hum. Genet., *58:* 1347–1363, 1996.
11. Kong, A., and Cox, N. J. Allele-sharing models: LOD scores and accurate linkage tests. Am. J. Hum. Genet., *61:* 1179–1188, 1997.
12. Ostrander, E. A., and Stanford, J. L. Genetics of prostate cancer: too many loci, too few genes. Am. J. Hum. Genet., *67:* 1367–1375, 2000.

Genetic Epidemiology 21(Suppl 1): S429–S434 (2001)

Exploiting Linkage Disequilibrium in Population Isolates

Robert Culverhouse, Jennifer Lin, Kuang-Yu Liu, and Brian K. Suarez

Departments of Psychiatry (R.C., J.L., K.-Y.L., B.K.S.) and Genetics (B.K.S.), Washington University School of Medicine, St. Louis, Missouri

A comparison of haplotype frequencies between unrelated cases and controls from population isolates identifies a strong signal for the presence/absence of the discrete phenotype in an interval on chromosome 6 bounded by markers 34 and 35. We define a dissimilarity index, D, which is sensitive to differences in allele distributions and differences in the patterns of linkage disequilibrium between cases and controls. We describe two statistical methods to utilize D: a method appropriate for a single moderately sized sample and a sequential approach appropriate for multiple small independent samples. © 2001Wiley-Liss, Inc.

Key words: case/control, haplotypes, linkage disequilibrium

INTRODUCTION

Linkage disequilibrium (LD) in a population refers to the nonrandom association of alleles at two (or more) loci on the same chromosome. The ultimate source of LD is the occurrence of a mutation that results in complete disequilibrium of the mutant allele with alleles that happen to occupy the same chromosome. Over time, disequilibrium decays as a consequence of recombination.

The strategy of fine mapping a chromosomal region that contains a disease susceptibility (DS) locus using linkage disequilibrium has intuitive appeal in three settings: (1) a population begun by a few founders; (2) a population that has experienced a severe bottleneck; and (3) an admixed population. The approach has proved successful in detecting rare, highly penetrant genes [Hästabacka et al., 1992; Sulisalo et al., 1994].

It is not obvious, however, that LD in a genetic isolate will be useful for detecting DS loci for common phenotypes. Clearly, detectability of a DS locus depends on many factors including the size of its phenotypic effect, disease heterogeneity in the isolate,

Address reprint requests to Dr. Robert Culverhouse, Washington University School of Medicine, Box 8134, Dept. of Psychiatry, 660 South Euclid, St. Louis, MO 63110.

the number of founding copies of the disease allele(s), and whether the DS alleles occur on common or rare haplotypes.

Most measures used to quantify the degree of disequilibrium simply compare the allele frequencies in cases and controls at a single marker [Devlin and Risch, 1995]. In this study we investigate whether a comparison of haplotypes that capitalizes on the frequency differences between cases and controls can successfully map susceptibility loci for a common complex phenotype.

METHODS

Each of the 50 population isolates simulated for Genetic Analysis Workshop (GAW) 12 contained 165 nominally unrelated founders (i.e., individuals for whom no parents are specified, but who are expected to be distantly related to one another through common ancestors). The isolates were founded by approximately 100 individuals about 20 generations ago. For each isolate, the sample of cases consisted of all affected individuals from these 165 "unrelated founders." The average number of affected individuals per isolate was 56 (range 42 to 69). An equal number of unaffected controls was selected from the remaining founders. Because the prevalence of the disease increases with age, we chose as controls the oldest unaffected individuals to minimize the number who might be carrying DS alleles.

Prior to undertaking any comparisons between cases and controls, we used the ASSOCIATE program [Ott, 1985] to examine the evidence for allelic association for all pairs of adjacent markers as a function of the genetic distance separating them in the first isolate (Figure 1A) and in the first general population replicate (Figure 1B). All 165 founders were used for this analysis. The evidence for disequilibrium is striking in the isolates (highly significant for most pairs of marker loci, even when separated by several centimorgans), suggesting that the isolates were not established by a random sample of unrelated founders from the general population 20 generations ago. The presence of such extensive of disequilibrium also suggests that a strategy that simply searches for regions with increased disequilibrium in cases versus controls is unlikely to succeed.

For each pair of adjacent markers, we obtained maximum likelihood estimates of the haplotype frequencies separately for cases and controls after 50 iterations of the EM algorithm from the ASSOCIATE program. A test of 100 adjacent marker pairs revealed

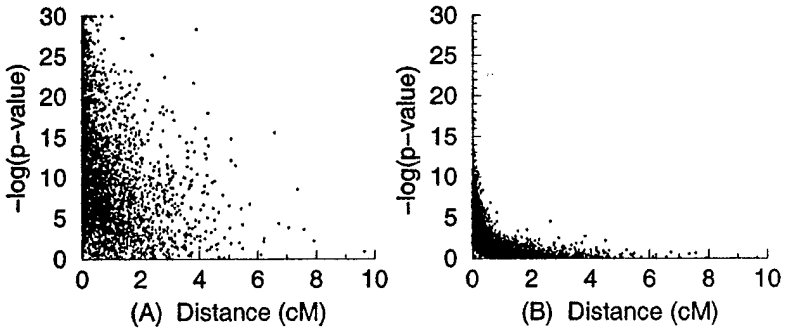

Fig. 1. Allelic association p-values for adjacent markers versus intermarker distance in isolate population replicate 1 (A), and in general population replicate 1 (B).

that 50 iterations was sufficient to attain covergence at a tolerance of 0.0001 between successive log likelihoods. Since the haplotype estimation procedure used in ASSOCIATE assumes that marker genotypes are in Hardy-Weinberg equilibrium (HWE), we tested this assumption for all markers on haplotypes that result in significant dissimilarity between cases and controls. Our measure of the dissimilarity between the haplotype distributions of cases and controls was defined as follows: Let m be the number of alleles at marker 1 and n the number of alleles at marker 2. Then

$$D = \sqrt{\sum_{ij=1}^{m,n} (a_{ij} - u_{ij})^2} \quad ,$$

where a_{ij} and u_{ij} are the inferred proportion of haplotypes consisting of the i^{th} allele at marker 1 and the j^{th} allele at marker 2 in the cases and controls, respectively.

The dissimilarity index, D, is the norm of the vector difference between cases and controls. Without claiming optimality for this choice of D, we note that it has some appealing features. First, squaring the elements gives more weight to a single large difference than to several small differences. This is expected to reduce the accumulation of noise in large tables. Second, the D statistic is sensitive both to differences in allele distributions and to differences in the pattern of LD between cases and controls.

Because the number of cases in any isolate is so small (average 56), to obtain a sample of sufficient power required using cases and controls from more than one isolate. We took two approaches to this problem: a sequential testing of the small individual isolates, appropriate if the isolates are independent (Figure 2, Model I), and a fixed sample size test, appropriate if the replicates were all derived from a single population allowing unrelated cases and controls to be combined from multiple isolates (Figure 2, Model II).

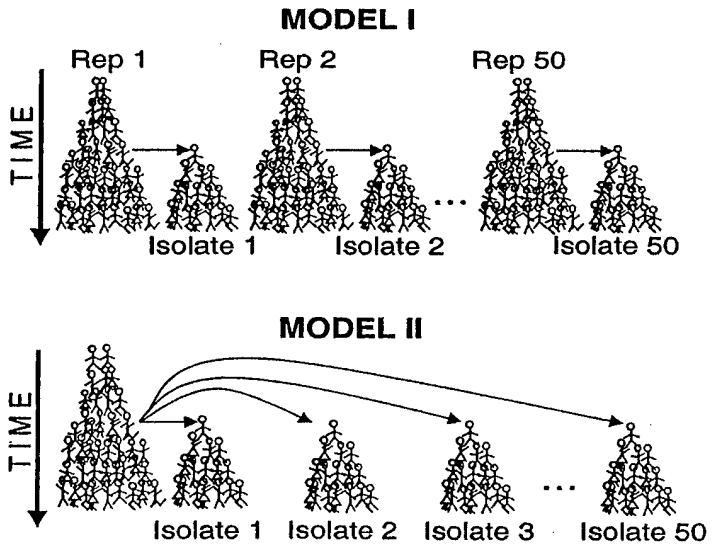

Fig. 2. Models of isolate formation. Model I postulates that each isolate is independent of all other isolates. Model II postulates that all isolates are derived from the same parental population.

If the isolates were constructed from independently generated populations, it would be inappropriate to pool samples because different marker alleles would be associated with the disease alleles in different isolates. For this reason we used a sequential test that terminates once sufficient evidence for or against the alternative hypothesis accumulated. The sequential test we used was based on the ranked values of D within each chromosome. The decision procedure is described in Dixon and Massey [1969].

For a given marker pair M, the following hypotheses are evaluated sequentially:

H_0: The rank order of the D value for adjacent marker pair M is strictly random.

H_1: The D values for adjacent marker pair M will rank in the upper decile for at least 20% of the trials.

Dividing the rank by the number of marker pairs on the chromosome produces a value approximately equal to the probability (under H_0) of obtaining this rank or higher. For the k^{th} trial we will refer to this probability as p_k. Under the null hypothesis, the distribution of these probabilities is approximately uniform on [0,1]. To compute p_{oj}, the probability of the first j observations (trials) under H_0, we used the Fisher [1932] statistic ($-2\sum\ln(p_k)\sim\chi^2$). The values p_{1j} (defined analogously) were computed in the following manner: a trial was considered a success if the rank fell in the upper decile and a failure otherwise. The likelihood of having at least as many successes as found in a particular sequence of j trials, p_{1j}, was computed using a binomial distribution with a parameter value of 0.2. Using twice the expectation under H_0 to test for elevated ranks (H_1) is arbitrary and the effect of using other thresholds was not examined.

The ranking process involves many comparisons on each chromosome. For this reason, to achieve a significance level of 0.01 on a chromosome, we used an α corrected for the number of adjacent marker pairs on the chromosome. For a chromosome with N marker pairs, we used the correction $\alpha(C) = \alpha/N$, which is conservative even for dependent tests.

To confirm a detection, we reinitiated the process whenever there were at least 25 isolates (½ of the total) remaining. All analyses using the sequential method were conducted without knowledge of the generating model.

The above method is appropriate for small independent samples. However, if the small samples are all derived from a single population, the sequential method will result in an inflation of the number of false positive signals, because artifacts of stochastic variation in the population would be repeated across the replicate samples.

In such a setting, illustrated in Model II (Figure 2), multiple replicates can be pooled to form a single, moderately sized epidemiological sample of "unrelated" cases and controls, assuming that all isolates are of similar age and that their antiquity is not too great. Accordingly, after learning that the isolate replicates were all derived from a single population, we pooled "unrelated" cases and controls from the first 10 replicates. This resulted in a sample of 575 cases and an equal number of controls (approximately the same total number of genotypes as in a single isolate). The significance of a "signal" in this pooled sample was evaluated under the hypothesis that the D values for markers unlinked to a DS locus would have a distribution approximated by the normal $N(D_m,D_{var})$, where D_m is the mean of the D values in the sample and D_{var} is the sample variance. A subsequent analysis of the distribution of D values on chromosomes containing no DS loci showed that D is normally distributed (data not shown). All analyses using the pooled sample were made with knowledge of the generating model.

RESULTS

Under the sequential testing method, only two adjacent marker pairs reached chromosome-wide 0.01 significance in the first 25 replicates and were confirmed in the remaining replicates: the haplotypes consisting of markers 34-35 on chromosome 6, and haplotypes consisting of markers 35-36 on chromosome 11. Of the 2,833 tested adjacent marker pairs, 1,668 (58.9%) reached the lower critical boundary (i.e., H_0 was accepted and sampling stopped) on or before the 50th replicate. Of the remaining marker pairs, 2 reached the upper boundary in ≤ 25 replicates, but did not repeat, 11 reached the upper boundary after more than 25 replicates, and 1,150 did not reach either boundary in 50 replicates. Information on the two "confirmed" marker pairs is given in Table I.

Under the single sample method, only the value of D at marker pair 34-35 and its neighboring haplotypes on chromosome 6 were significant at $p < 0.01$ after Bonferroni correction. Because this method involves the distribution of D values across the genome, rather than the chromosomal rank of D, the multiple comparison correction is made for all 2,833 adjacent marker pairs tested. The corrected p-value for marker pair 34-35 is 1.8×10^{-28}. The dramatic difference between the height of the signal on chromosome 6 and the highest non-chromosome 6 peak is shown in Figure 3. Analyses of markers 34 and 35 on chromosome 6 and markers 35 and 36 on chromosome 11 show that none deviate significantly from HWE in either the cases or controls. It is noteworthy that the false positive signal detected on chromosome 11 with the sequential testing strategy (Table I) is not detected in the pooled sample.

DISCUSSION

Susceptibility loci for common diseases are proving difficult to map by linkage analysis because their effects are relatively small and because common diseases are heterogeneous. In principle, techniques that exploit the presence of linkage disequilibrium should have increased power to detect DS loci, but the choice of an appropriate study population is critical [Wright et al., 1999]. In many ways the isolates simulated for GAW12 are less than ideal for a case/control study because there are so few "unrelated" persons per isolate and because the 5% immigration rate per generation stretches the notion of "isolate."

Under the (mistaken) hypothesis that the GAW12 isolates were independently generated (corresponding to Model I of Figure 2), a sequential testing strategy that preserved the informational integrity of each isolate was developed. In the GAW12 isolates this strategy correctly identified the location of MG6, but also gave rise to false detections.

TABLE I. Marker Pairs That Reached Chromosome-Wide Significance in ≤ 25 Replicates, and Were Confirmed in the Remaining Replicates ($\alpha = 0.01$, $\beta = 0.05$)

Chromosome	Marker pair	$\alpha(C)$	Reps to decision	Reps to confirm
6	34–35	6.7×10^{-5}	8	20
11	35–36	8.7×10^{-5}	25	18

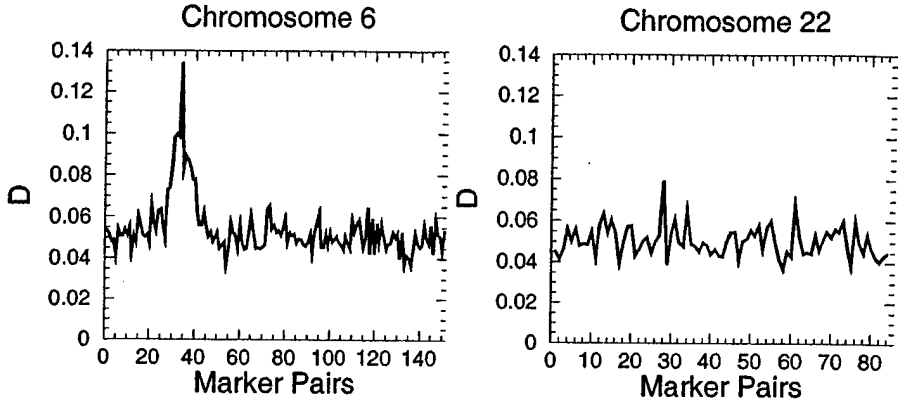

Fig 3. Left, distribution of D values for chromosome 6 using the first 575 cases and 575 controls. Right, distribution of D values for chromosome 22 from first 575 cases and 575 controls. The D value for marker pair 28-29 on chromosome 22 is the highest nonchromosome 6 signal.

In contrast, using data pooled from multiple isolates to produce a single epidemiological sample is appropriate whenever multiple isolates have recently budded from a single parental population (as illustrated in Model II of Figure 2). This model more nearly approximates how the GAW12 isolates were generated. Using a single sample of 575 unrelated cases and 575 unrelated controls from the first 10 isolates, (a sample of approximately the same number of genotypes as in a single replicate) the position of the DS locus that directly affects the presence/absence of the discrete phenotype is precisely located, and no false positives were detected after Bonferroni correction. While the pooling strategy was necessitated by the small number of cases in each isolate, the natural setting for the single sample approach is a single isolated population.

ACKNOWLEDGMENTS

This work was supported, in part, by USPHS grants MH14677 and MH31302, US Army grant DAMD17-00-10108, and an award from the Urological Research Foundation.

REFERENCES

Devlin B, Risch N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311-22.

Dixon WJ, Massey FJ. 1969. Introduction to statistical analysis. New York: McGraw Hill.

Fisher RA. 1932. Statistical Methods for Research Workers, 4th ed. London: Oliver & Boyd.

Hästabacka J, de la Chappelle A, Kaitila I, et al. 1992. Linkage disequilibrium mapping in isolated founder populations: Diastrophic dysplasia in Finland. Nat Genet 2:204-11.

Ott J. 1985. A chi-square test to distinguish allelic association from other causes of phenotypic association between two loci. Genet Epidemiol 2:79-84.

Sulisalo T, Klockars J, Mäkitie O, et al. 1994. High-resolution linkage-disequilibrium mapping of the cartilage-hair hypoplasia gene. Am J Hum Genet 55:937-45.

Wright AF, Carothers AD, Pirastu M. 1999. Population choice in mapping genes for complex diseases. Nat Genet 23:397-304.

A Perspective on Epistasis: Limits of Models Displaying No Main Effect

Robert Culverhouse[1], Brian K. Suarez[1,2], Jennifer Lin[1], Theodore Reich[1,2]

[1]*Department of Psychiatry and* [2]*Department of Genetics*

Washington University School of Medicine, St. Louis, MO

Running Title: Epistatic Models With No Main Effect

Address for correspondence:

Robert Culverhouse

Department of Psychiatry

Washington University School of Medicine

660 South Euclid

St. Louis, MO 63116-1026

Telephone: (314) 362-8077

Fax: (314) 362-4247

e-mail: rob@frodo.wustl.edu

key words: epistasis, linkage, association, multilocus analysis, complex diseases

Summary

The completion of a draft sequence of the human genome and the promise of rapid SNP genotyping technologies have resulted in a call for the abandonment of linkage studies in favor of genome scans for association. However, there exists a large class of genetic models for which this approach will fail: purely epistatic models with no additive or dominance variation at any of the susceptibility loci. As a result, traditional association methods (such as case/control, measured genotype, and TDT) will have no power if the loci are examined individually.

In this paper we examine this class of models, delimiting the range of genetic determination and recurrence risks for 2-, 3-, and 4-locus purely epistatic models. Our study reveals that these models, while giving rise to no additive or dominance variation, do give rise to increased allele sharing between affected sibs. This in turn implies that a genome scan for linkage examining loci serially might detect the susceptibility loci. We also discuss some natural multilocus extensions of single locus analysis methods, including a conditional form of the TDT.

Introduction

A quarter century ago the advent of recombinant DNA technology spurred what can arguably be called the "Golden Age of Human Linkage Studies." The availability of a large number of new DNA markers that could be typed directly (e.g. RFLPs, VNTRs, microsatellites), coupled with major advances in computer software and hardware meant that mapping genes that are individually sufficient to cause human disease rested on little more than collecting an adequate sample of segregating families.

The approach is straightforward. A genome scan using 300-400 short sequence tandem repeat markers reveals a single linkage signal. Additional markers are added to saturate this chromosomal region. Subsequent identification of recombinants delimits the genomic segment containing the disease-causing gene. A search is then initiated to identify all variants in the region (insertions, deletions, SNPs, repeat polymorphisms, etc.) and these, in turn, are tested for cosegregation with the disease, usually in the same families that provided the original linkage signal. Once the genetic lesion is identified, funtional analysis is used to clarify how the lesion alters disease susceptibility.

This idealized scenario is not intended to trivialize the hard work needed to successfully complete each step. Ten years, for instance, elapsed between mapping the locus responsible for Huntington's Disease to chromosome 4p (Gusella et al. 1983) and the subsequent identification of the expanded CAG repeat in the gene's first exon (Huntington's Disease Collaborative Research Group 1993). Nonetheless, a litany of disease genes that have been identified during the last two decades is ample testimony to this strategy's success.

But what about complex diseases? Is it reasonable to suppose that an approach that *must* succeed in identifying fully penetrant Mendelian genes will also succeed for complex diseases? Most gene sleuths would answer "yes" but add the caveat that since recombinants cannot be identified unambiguously, ancillary approaches are needed as well. Three approaches that often accompany linkage studies of complex diseases are: family-based association studies (Spielman et al. 1993), case/control studies (Woolf 1955), and measured genotype studies (Boerwinkle et al 1986). All of these ancillary approaches tacitly assume that allelic variation

in or around a *particular* susceptibility locus makes a measurable difference in the phenotype. The reasonableness of this assumption is so obvious that it is rarely explicitly stated.

If the advances wrought a generation ago by recombinant DNA technology can be said to have revolutionized genetic research, we believe that the field is poised to experience an even greater revolution in the near future. Against the backdrop provided by the completion of a draft sequence of the human genome, it is reasonable to expect that both single-nucleotide polymorphism (SNP) and expression microarray technologies will forever change how research in human genetics is pursued. Moreover, we believe these new technologies will make the three ancillary strategies mentioned above all the more attractive to researchers – even to the point of supplanting traditional linkage analysis – for at least two reasons. First, it has been argued that linkage methods simply do not have the power to detect the small signals that can be expected under some disease models (Risch and Merikangas 1996). Second, even when appropriate power can be obtained, multiplex pedigrees are substantially more expensive to gather than either a series of unrelated patients (for measured genotype analyses), a series of patients and "matched" unaffected subjects (for comparison with patients in case/control studies), or parents and a single affected offspring (for family-based association studies).

The complex relationship between genotype and phenotype, however, may ultimately prove to be inadequately described by simply summing the modest effects from several contributing loci. Instead, the relationship may, as Sewell Wright argued (1923), depend in a fundamental way on epistasis (the interaction between loci) and genotype-by-environment interaction. Indeed, it has been argued that epistatic interactions are a nearly universal component of the architecture of most common traits. Templeton (2000), for instance, lists a number of phenotypes where epistasis plays a large role. An example in insects is the abnormal abdomen phenotype in *Drosophila mercatorum* (DeSalle and Templeton 1986; Hollocher et al. 1992; Hollocher and Templeton 1994). In humans, variation in triglyceride levels can be explained, in part, by two sets of interactions: between ApoB and ApoE in females and between the ApoAI/CIII/AIV complex and LDLR in males (Nelson et al.

2001). Even the seemingly "simple" Mendelian trait of sickle-cell anemia is revealed to be greatly modified by epistatic interactions. Sickle-cell individuals who are homozygous for two polymorphisms near the $G\gamma$ locus (leading to the persistence of fetal hemoglobin) have only mild clinical symptoms (Odenheimer et al. 1983; Sing et al. 1985; el-Hazmi et al. 1992). Other human diseases which have recently been reported to exibit epistatic interactions are Alzheimer's disease (Zubenko et al. 2001) and breast cancer (Ritchie et al. 2001).

The main reason most studies of complex human phenotypes fail to find evidence for epistatic interactions may simply be that commonly used designs and analytic methods inherently minimize or exclude the possibility of epistasis (Frankel and Schork 1996). Because the investment in new study designs and analytic methods may be high, we decided to examine the extent to which *purely* epistatic interactions (interactions between loci that do not display any single locus effects) could account for phenotype.

In this paper, we explore a class of transmission models where each contributing susceptibility locus has no phenotypic effect detectable by any of the three ancillary approaches listed above. For these epistatic models, the proper "unit of analysis" is not the allelic variation at a single locus, but the multilocus genotype. We show that, in spite of the undetectability of the contributing loci by the three ancillary approaches, these models can result in both high heritability and substantial increases in recurrence risk to a proband's relatives. In addition, we show that the within-family increase in allele sharing engendered by these models can be sufficient to allow the detection of the contributing loci by linkage.

Models

Consider a dichotomous qualitative trait (e.g. "affected" vs "unaffected" phenotype) determined by L biallelic loci. We examine the extent to which affection status can be genetically determined ("broad sense" heritability) in models for which the marginal penetrance for each of the three genotypes is equal to the population prevalence of the disease (K) for each of the contributing loci. In what follows, we assume all disease susceptibility loci are in Hardy-Weinberg equilibrium and alleles at different susceptibility loci are in linkage

equilibrium.

Two-Locus Models

Let the alleles from locus A be denoted A and a, while B and b denote the alleles from locus B. Let AA be genotype 1 at locus A, Aa be genotype 2, and aa be genotype 3. The genotypes at locus B are defined in the corresponding manner. Let p_A be the allele frequency of A and p_B the frequency of B. Let f_{ij} be the disease penetrance for the genotype consisting of genotype i at locus A and genotype j at locus B. M_{Ai} denotes the marginal penetrance for genotype i at locus A, while M_{Bj} denotes the marginal penetrance for genotype j at locus B.

The relationship between these variables is given by the following formulae:

$$\begin{aligned}
M_{A1} &= p_B^2 f_{11} + 2p_B(1-p_B)f_{12} + (1-p_B)^2 f_{13} \\
M_{A2} &= p_B^2 f_{21} + 2p_B(1-p_B)f_{22} + (1-p_B)^2 f_{23} \\
M_{A3} &= p_B^2 f_{31} + 2p_B(1-p_B)f_{32} + (1-p_B)^2 f_{33} \\
M_{B1} &= p_A^2 f_{11} + 2p_A(1-p_A)f_{21} + (1-p_A)^2 f_{31} \\
M_{B2} &= p_A^2 f_{12} + 2p_A(1-p_A)f_{22} + (1-p_A)^2 f_{32} \\
M_{B3} &= p_A^2 f_{13} + 2p_A(1-p_A)f_{23} + (1-p_A)^2 f_{33} \\
K &= p_A^2 M_{A1} + 2p_A(1-p_A)M_{A2} + (1-p_A)^2 M_{A3} \\
K &= p_B^2 M_{B1} + 2p_B(1-p_B)M_{B2} + (1-p_B)^2 M_{B3}
\end{aligned} \tag{1}$$

and is commonly represented by the penetrance table seen in Table 1. For all of the genetic variation to be epistatic, a two-locus model must also satisfy

$$M_{Ai} = M_{Bi} = K \qquad \forall i \in \{1,2,3\} \tag{2}$$

[Table 1 about here]

To explore the space of penetrance models that satisfy (1) and (2), we varied K, p_A, and p_B. For each set of parameter values, we searched for a penetrance model which would maximize the proportion of variation attributable to genotype. The total variance of the dichotomous phenotype in the population is $V_T(K) = K(1-K)$. For models satisfying

formulae (1) and (2), all of the variation attributable to genotype is epistatic. This variation, V_I, is given by the following formula:

$$\begin{aligned} V_I(\vec{f}, K, p_A, p_B) &= p_A^2 p_B^2 (f_{11} - K)^2 + 2p_A^2 p_B (1 - p_B)(f_{12} - K)^2 \\ &+ p_A^2 (1 - p_B)^2 (f_{13} - K)^2 + 2p_A (1 - p_A) p_B^2 (f_{21} - K)^2 \\ &+ 4p_A (1 - p_A) p_B (1 - p_B)(f_{22} - K)^2 + 2p_A (1 - p_A)(1 - p_B)^2 (f_{23} - K)^2 \\ &+ (1 - p_A)^2 p_B^2 (f_{31} - K)^2 + 2(1 - p_A)^2 p_B (1 - p_B)(f_{32} - K)^2 \\ &+ (1 - p_A)^2 (1 - p_B)^2 (f_{33} - K)^2 \end{aligned} \tag{3}$$

Thus, for fixed K, p_A, and p_B, maximizing the broad heritability ($h^2 = V_I/V_T$) under constraint (2) is equivalent to maximizing V_I.

Constraints (1) imply that, for fixed K, p_A, and p_B, the remaining five penetrances can be written as linear combinations of the four corner penetrances, (f_{11},f_{13},f_{31}, and f_{33}). Therefore, the set of f_{ij} satisfying (1) and (2) forms a 4-dimensional polyhedral subset of the 9-dimensional unit hypercube, $\prod_{i=1}^{9}[0, 1]$. Our goal is to maximize V_I over this polyhedron. A linear transformation converts this problem to the problem of maximizing the distance between a fixed point in the interior of the polyhedron and other points in the polyhedron. Therefore, one of the vertices of the polyhedron must correspond to a model generating the maximum heritability. Determining the vertices of the polyhedron is a linear algebra problem which can, in theory, be solved explicitly for any fixed values of our parameters. Because of the number of constraints involved, we used the cdd+ program (Fukuda 1999) which implements the "Double Description Method" (Motzkin et al. 1953) to find the verticies of the polyhedra of 2-locus purely epistatic models.

Table 2 lists the maximum heritabilities for various combinations of K, p_A, and p_B. For each K, the greatest heritability was found when $p_A = p_B = 0.5$

[Table 2 about here]

An alternative visualization tool is to vary K while keeping p_A and p_B constant. Using cdd+, we were able to parametrize the verticies of the space of 2-locus purely epistatic models

when $p_A = p_B = 0.5$ in terms of K. The range of K examined was $(0, 1/2]$ because the maximum heritabilities are necessarily symmetric about $K = 1/2$. We found that there are 7 verticies when $K \in (0, 1/4]$ and 25 verticies when $K \in (1/4, 1/2]$. The curves corresponding to the maximum heritabilities are described by the following formulae:

$$V_I(K) = \begin{cases} 2K^2 & \text{if } K \in (0, 1/4] \\ 2K^2 - K + \frac{1}{4} & \text{if } K \in [1/4, 1/2] \end{cases} \quad (4)$$

$$h^2 = h^2_{max}(K) = \begin{cases} \frac{2K}{1-K} & \text{if } K \in (0, 1/4] \\ \frac{2K^2-K+1/4}{K(1-K)} & \text{if } K \in [1/4, 1/2] \end{cases} \quad (5)$$

These maxima can be achieved using the penetrances in Table 3 for $K \in [1/4, 1/2]$ and the values given in Table 4 for $K \in (0, 1/4]$.

[Tables 3 and 4 about here]

A plot of population prevalence versus maximum heritability is shown in Figure 1. This graph also contains a plot of K versus total variance and of K versus maximum epistatic variance.

[Figure 1 about here]

Figure 1 illustrates that there exist two-locus models with no marginal genotypic effect at either locus but in which genotype nonetheless accounts for a large portion of the population variance. Although the heritability can be high in these models, the constraints that eliminate any marginal gene effects keep the recurrence risks modest. The relative risks to offspring (λ_o) and to sibs (λ_{sib}) of an affected individual are given by the following formulae:

$$\lambda_o = \begin{cases} 1.25 & \text{if } K \in (0, 1/4] \\ 1.25 - \frac{1}{4}(\frac{K-1/4}{K^2}) & \text{if } K \in [1/4, 1/2] \end{cases} \quad (6)$$

$$\lambda_{sib} = \begin{cases} 1.3125 & \text{if } K \in (0, 1/4] \\ 1.3125 - \frac{1}{4}(\frac{K-1/4}{K^2}) & \text{if } K \in [1/4, 1/2] \end{cases} \quad (7)$$

Three-Locus Models

Although the verticies of the polyhedra of purely-epistatic models can, in theory, always be found, in practice it proved computationally impractical for many three-locus parameter sets. These often involve several thousand vertices. We found that a slight perturbation in K could lead to a 20,000-fold increase in computing time when using the Fukuda (1999) program.

For this reason, we estimated the maxima for three-locus models using the non-linear maximization methods implemented in the SAS (1995) procedure PROC NLP. Because this method estimates local rather than global maxima, we used 1000 random seeds for each parameter set, choosing the highest resulting epistatic variance as our approximation of the true maximum. We verified this approach on two-locus models, finding the true maximum at each of 100 points.

Three-locus models produce a dramatic increase in the maximum proportion of variation explainable by genotype, as can be seen in Figure 2. For the three-locus case we have again found specific models that closely fit the numerically derived maxima plotted as dots in Figure 2. The curves generated by these models are drawn as a line beneath the dots. The fact that we could find models with V_I at least as high as the iterative estimates for each K indicates that rounding errors from SAS are not likely to cause substantial overestimation of the maximum heritability. In fact, for a few points near $K = 0.4$ and $K = 0.46$ the empirical estimate slightly underestimated the true maximum.

[Figure 2 about here]

Formula (8) describes the estimated maximum epistatic variance curve for all models involving three loci. It was derived using specific models we found using our maximization search method. The third piece of the curve ($V_I = \frac{11}{2}K^2$ for $K \in [\frac{2-\sqrt{1/2}}{14}, \frac{1}{8}]$) has a different form than the other pieces and has anomolous limits. Nonetheless, checks of the verticies of the solution polyhedra for several values of K in and around this region confirm the estimated values.

$$V_I = \begin{cases} 9K^2 & \text{if } K \in (0, \frac{1}{16}] \\ 9K^2 - K + \frac{1}{16} & \text{if } K \in [\frac{1}{16}, \frac{2-\sqrt{1/2}}{14}] \\ \frac{11}{2}K^2 & \text{if } K \in [\frac{2-\sqrt{1/2}}{14}, \frac{1}{8}] \\ 9K^2 - \frac{9}{4}K + \frac{29}{128} & \text{if } K \in [\frac{1}{8}, \frac{5}{32}] \\ 9K^2 - 2K + \frac{3}{16} & \text{if } K \in [\frac{5}{32}, \frac{3}{16}] \\ 9K^2 - 3K + \frac{3}{8} & \text{if } K \in [\frac{3}{16}, \frac{1}{4}] \\ 9K^2 - 5K + \frac{7}{8} & \text{if } K \in [\frac{1}{4}, \frac{9}{32}] \\ 9K^2 - \frac{17}{4}K + \frac{85}{128} & \text{if } K \in [\frac{9}{32}, \frac{5}{16}] \\ 9K^2 - \frac{25}{4}K + \frac{165}{128} & \text{if } K \in [\frac{5}{16}, \frac{11}{32}] \\ 9K^2 - \frac{11}{2}K + \frac{33}{32} & \text{if } K \in [\frac{11}{32}, \frac{3}{8}] \\ 9K^2 - \frac{15}{2}K + \frac{57}{32} & \text{if } K \in [\frac{3}{8}, \frac{13}{32}] \\ 9K^2 - \frac{27}{4}K + \frac{189}{128} & \text{if } K \in [\frac{13}{32}, \frac{7}{16}] \\ 9K^2 - \frac{35}{4}K + \frac{301}{128} & \text{if } K \in [\frac{7}{16}, \frac{15}{32}] \\ 9K^2 - 8K + 2 & \text{if } K \in [\frac{15}{32}, \frac{1}{2}] \end{cases} \quad (8)$$

The maximum possible heritability in models with no single-locus additive or dominance variance increases dramatically from 2-locus to 3-locus models. Two-locus models require a disease prevalence above 47% for heritability to reach 90% while 3-locus models can be completely genetic for prevalences as low as 25%. Furthermore, for K between 0.05 to 0.10, a range that includes the prevalances of many complex diseases, 3-locus models can generate heritabilities from 35% to 55%. In contrast, purely epistatic 2-locus models can only generate 10% to 22% heritability.

Three-locus models can also give rise to higher relative risks than are possible in corresponding two-locus models. Three-locus penetrance models maximizing heritability at the low end of disease prevalence ($K \in (0, 1/16]$) are parameterized in Table 5. These models correspond to a $\lambda_{sib} = 2.125$. In contrast, the highest λ_{sib} possible for 2-locus epistatic models is 1.3125.

[Table 5 and Figure 3 about here]

Because none of the alleles in these models have any marginal effect on disease susceptibility, the disease would not cause selection pressure on allele frequency at any of the loci. Nonetheless, genetic drift, mutation, and selection pressure from factors other than the disease in question are likely to perturb allele frequencies from 50%/50%. Figure 3 illustrates the effect unbalanced allele frequencies have on these models. In the figure, p_i denotes the frequency of the less common allele at locus i. (This should not be interpreted as the "disease allele" frequency. In these models, there are no "disease alleles," only "disease genotypes.") While a smaller proportion of the total variance is attributable to genotype in models with these unbalanced allele frequencies (40%/60%, 20%/80%, and 20%/80%) than when all alleles are equally frequent, a sizable portion of the variation can still be explained by genotype.

Four-Locus Models

Figure 4 illustrates the estimated maximum heritabilities possible for models involving four interacting loci. The maximum heritability remains over 90% for prevalences above 12% and maximum heritability is not much less than 50% unless prevalence is below 2%. Furthermore, for prevalences between 0% and 2%, the maximum heritability possible with four loci is approximately four times as high as for models involving only three loci. Some of the jaggedness of the figure may be attributable to the fact that points are plotted in increments of 0.0025 for K and that only 1000 iterations were used for each point. We have observed that using too few iterations can considerably underestimate the maximum heritability.

[Figure 4 about here]

As before, the addition of a locus corresponds to an increase in recurrence risk to relatives. At the low end of disease prevalence ($K < 0.0156$), λ_{sib} can reach 2.609 (compared to $\lambda_{sib} = 2.125$ for 3-locus models and $\lambda_{sib} = 1.3125$ for 2-locus models).

Four-locus models also appear more robust to perturbation of allele frequencies than 3-locus models. Figure 5, for instance, displays results from 4-locus models with frequencies

of the less common alleles being 20% at locus A, 30% at locus B, 40% at locus C, and 50% at locus D. Maximum heritability remains over 95% for disease prevalences above 16% and does not fall below 50% unless the disease prevalence is below 4%.

[Figure 5 about here]

Detecting epistatically interacting loci

Association

By design, the models described here have equal marginal penetrances for all single-locus genotypes. Because of this, it is obvious that a qualitative measured genotype analysis of a single locus (analogous to the quantitative measured genotype of Boerwinkle et al. (1986)) could not detect any of the contributing loci. Furthermore, the equality of the marginal penetrances implies that cases and controls would have identical allele distributions at the contributing loci. Thus, a case-control study examining one locus at a time would also fail to detect the contributing loci.

A TDT (Spielman et al. 1993) study examining a single locus at a time would also fail to detect the contributing loci, but the reasons may seem less obvious. Within particular families, heterozygous parents would preferentially transmit the allele A to their affected offspring. However, in a balancing proportion of families, heterozygous parents would preferentially transmit the allele a to affected offspring. The TDT, in common with other association analyses, keeps track of the particular "at risk" allele that is differentially present in affected individuals or is preferentially transmitted to affected offspring. Thus, families segregating allele A will "cancel out" the evidence from families segregating allele a. As a result, under these purely epistatic models, the TDT statistic at the contributing loci will be equivalent to those from "neutral" loci.

Since it is impossible to detect these loci using a locus-by-locus genome scan for association, one might consider a scan assessing two or more loci at a time. Estimates of the number of SNPs required for a whole genome scan range from as many as 500,000 (Kruglyak 1999) to

as few as 30,000 (Collins et al. 1999). To examine all 2-way interactions for even the smaller number would require approximately 450 million tests. To examine all 3-way interactions would require approximately 4.5 trillion tests. These are non-trivial computational tasks, not to mention the statistical problem of correcting for multiple tests.

The fact that the number of tests involved in examining interactions grows as a polynomial in the number of loci suggests that successful analyses of interactions will depend on a method of selecting a limited number of candidate loci for consideration. Fortunately, for the class of models discussed in this paper, linkage analysis is often capable of detecting increased allele sharing at loci that epistatically contribute to the affected phenotype.

Linkage

Although the purely epistatic models discussed here do not give rise to different allele frequencies in cases and controls, they do give rise to excess allele sharing among affected sibs. Because these epistatic models have no "disease alleles" (only "disease genotypes"), the allele that is shared excessively among affected sibs varies depending on the mating type of the parents. However, in contrast to association analyses, if half of the families in a linkage analysis show increased sharing for allele A at locus A and the other half show increased sharing for allele a, the linkage statistic at locus A for the combined sample is *higher* than in either subsample. Because the linkage statistic from each family is not tied to a specific allele, the evidence for linkage from families segregating for different alleles accumulates rather than cancels. Consider, for instance, a collection of affected sib pairs for a disease that conforms to the two-locus model of Table 4. Thirty-five of the possible 45 parental mating types are capable of segregating an affected child and 28 of these mating types give rise to sib pairs with increased allele sharing at locus A, locus B, or both. Indeed, at each locus the expected proportion of alleles shared identical-by-descent is 4/7 (calculations not shown) regardless of the value of K. Hence, regions containing both loci are detectable by linkage analysis, provided the sample size is adequate.

Analysis of Candidate Loci

Once a limited number of candidate loci are selected, it is feasible to examine the candidates for interactions. We note that the number of tests required to evaluate all 2-, 3-, and 4-way interactions for between 30 and 60 candidate loci has a range similar to the number of tests suggested for a single genome-wide scan for association using SNPs (Kruglyak 1999, Collins et al. 1999). Thus, although searching for 2-, 3-, 4-, or n-way interactions among *all* the markers in a genome scan would not be practicable, a candidate locus approach based on a genome scan for linkage may be.

Deriving appropriate and powerful methods to detect epistatic interactions remains a matter for further study. However, several straightforward methods are immediately available and some more elaborate methods are already in the literature.

Three elementary multilocus methods

Cases-only

The most straightforward multilocus analysis of cases-only data is a χ^2 test of independent segregation for the loci.

An analysis of data from the two-locus models described in Table 4, for instance, yields an expected test statistic $\geq 2N$, where N = the number of cases. This is a consequence of the fact that the expected value of the square of a random variable is at least as great as the square of the expected value of the variable. Under the null hypothesis of independent segregation, this statistic would be distributed as a χ^2 with 4 degrees of freedom.

Case/Control

A second approach is a multilocus case/control analysis. One method for doing this would be to compare the distribution of cases amongst the 3^L genotypes (L = the number of biallelic loci being simultaneously examined) with the distribution of controls.

Under this analysis, a sample of N cases and N unrelated controls draw from a population modeled by Table 4 will again yield an expected χ^2 statistic $\geq 2N$. However, the degrees of freedom under the null hypothesis are now 8. Moreover, compared to a cases-only strategy,

the inclusion of unrelated controls will add to the cost of genotyping. In addition, for diseases with a variable age-of-onset, the inclusion of controls who will eventually develop the disease will compromise power.

Conditional TDT

A third approach is a conditional TDT. For this, a sample of N trios is stratified by the genotype of the offspring at one (or more) of the candidate loci. A TDT analysis is then performed at another candidate locus for each stratum of the data. The p-values from the individual TDTs are then combined using Fisher's (1932) statistic ($S = -2\sum_{i=1}^{m} ln(p_i)$, where p_i is the p-value of the TDT corresponding to the i^{th} stratum of the data). Under the null hypothesis, S has a χ^2 distribution with m degrees of freedom, where m = the number of strata.

Consider, again, the model from Table 4. Conditioning on locus A, the trios with AA offspring would yield an expected χ^2 statistic = $N/4$ if a TDT analysis were performed at locus B. Trios with aa offspring yield the same expected χ^2 statistic. The trios with Aa offspring yield an expected χ^2 statistic of 1, independent of sample size. The Fisher statistic in this case would be tested against a χ^2 distribution with 3 degrees of freedom.

In our experience, the conditional TDT is less powerful than either the multilocus cases-only strategy or the multilocus case/control strategy. Of course, a conditional TDT approach will guard against unrecognized admixture in the sample. However, it also requires a three-fold increase in genotyping (2 parents and the affected offspring) compared to the cases-only strategy.

Other Methods

In addition to these simple methods, many other multilocus methods have been developed. Two-locus methods include a sib-pair analysis (Dizier and Clerget-Darpoux 1986), a two-locus lod score method (Lathrop and Ott 1990; Schork et al 1993), the marker-association-sequence χ^2 (MASC) method applied to two loci (Dizier et al 1994), and a two-locus version of the maximum lod score method (Cordell et al 1995).

Methods involving more than two loci are also under development. Nelson et al. (2001) used a combinatorial method for identifying multilocus genotypes contributing to variation in serum triglyceride levels. Ritchie et al. (2001) used a related data reduction technique to identify a four-locus risk factor for breast cancer.

Discussion

The models examined here are boundary cases. They represent extreme limits in terms of two parameters: the marginal deviation is at the lowest possible value and, given that constraint, the heritability is at the highest possible value.

Relaxing the maximum heritability condition leads to infinitely many models which still display no single-locus marginal effect but may appear more "natural" in that they have non-zero penetrances for all genotypes. Almost every purely epistatic model includes both incomplete penetrances and "phenocopies."

Relaxing the condition of zero single-locus marginal deviation results in a much larger class of models, but a less mathematically tractable class. Nonetheless, the heritabilities and λs found in the zero marginal deviation setting are useful in providing lower bounds for the maximum values possible when marginal effects are small. In particular, although the models specifically discussed here are boundary cases, they imply that seemingly "natural" models can account for most of the variation in disease even if all the single locus effects would pass undetected.

Although we have only examined 2-, 3-, and 4-locus models, the results lead us to two obvious extrapolations. First, we expect that, with a sufficient number of contributing loci, purely epistatic interactions could account for virtually all the variation in affection status for diseases with any prevalence. Second, models involving more loci could be associated with λs even greater than the 2.6 found for 4-locus models.

Of course, there are subclasses of purely epistatic models (providing no marginal evidence for the involvement of any single locus) for which, in addition, no two, three, or L-1 loci jointly give evidence of involvement in the disorder. This leads to the concern that even assessing

all 2-, 3-, and (L-1)-way interactions among candidate loci may be insufficient to detect the contributing loci.

This concern is ameliorated by the fact that such models are associated with lower heritabilities and much lower λ values than the models we have examined. The restriction on maximum heritabilities in these models is most easily seen by examining L-locus models for which no collection of L-1 loci show marginal deviations. The fact that all loci must pairwise satisfy the 2-locus constraints implies that the maximum heritability in this case is theoretically bounded above by the values from the two-locus model. In fact, a check of three-locus models with no deviation in the 2-way marginals (illustrated in Figure 6) shows that the true maximum heritability is even lower than the general upper bound.

[Figure 6 about here]

The fact that λ_{sib} will diminish exponentially with the number of simultaneously hidden loci is a simple consequence of the formula relating the covariance between siblings and the components of variation. For an L-locus system, this is given by the following formula (Kempthorne 1957):

$$Cov_{sib} = \frac{1}{2}V_A + \frac{1}{4}V_D + \frac{1}{4}V_{AA} + \frac{1}{8}V_{AD} + \frac{1}{16}V_{DD} + \ldots + \sum_{j+k=L} \left(\frac{1}{2}\right)^j \left(\frac{1}{4}\right)^k V_{A^j D^k}$$

For this reason, these models can only make a small contribution to λ.

Researchers of many complex diseases including non-insulin dependent diabetes mellitus, prostate cancer, and schizophrenia face the conundrum of moderately heritable diseases for which single-locus analyses have not accounted for the predicted genetic variance. The models discussed in this paper provide one possible explanation for this.

Had data been gathered for a disease that closely fit one of the epistatic models considered here, it is likely that the linkage signals from the contributing loci would have been rejected as "false positives." The impossibility of replicating or localizing the signals with single-locus follow-up association studies would make it very easy to reject the true signals.

These considerations lead us to believe that, in situations where heritability is moderate to high but single locus analyses do not account for the predicted genetic variance, it is worth pursuing a hypothesis of interacting loci near the linkage peaks. Even regions containing modest linkage signals may be good sources of candidate loci.

The epistatic models examined here were constructed so that none of the loci could be detected by single-locus case-control, measured genotype, or TDT studies. We found that a large fraction of the variation in affection status can be explained by such models for a wide range of population prevalences and allele frequencies. Less extreme, and therefore "more natural," models displaying small marginal effects can account for even more variation.

Since linkage analysis does not suffer from the drawbacks of single-locus association analyses for the class of epistatic models considered here, we conclude that it will continue to prove useful even when dense SNP maps are available and rapid genotyping becomes less costly.

Acknowledgements

We would like to thank Drs. Saurabh Ghosh and Anthony Hinrichs for their help. This work was supported, in part, by USPHS grants MH14677 and MH31302, US Army grant DAMD17-00-1-0108, and an award from the Urological Research Foundation.

References

Boerwinkle E, Chakraborty R, Sing CF (1986) The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. Ann Hum Genet 50:181-194

Collins A, Lonjou C, Morton NE (1999) Genetic epidemiology of single-nucleotide polymorphism. PNAS 96:15173-15177

Cordell HJ, Todd JA, Bennett ST, Kawaguchi Y, Farrall M (1995) Two-locus maximum lod score analysis of a multifactorial trait: joint consideration of IDDM2 and IDDM4 with IDDM1 in type 1 disease. Am J Hum Genet 57:920-934

DeSalle R and Templeton AR (1986) The molecular through ecological genetics of abnormal abdomen in *Drosophila mercatorum.* III. Tissue-specific differential replication of ribosomal genes modulates the abnormal abdomen phenotype in Drosophila mercatorum. Genetics 112:877-886

Dizier MH, Babron M-C, Clerget-Darpoux F (1994) Interactive effect of two candidate genes in a disease: extension of the marker-association-segregation χ^2 method. Am J Hum Genet 55:1042-1049

Dizier MH, Clerget-Darpoux F (1986) Two disease locus model: sib pair method using information on both HLA and Gm. Genet Epidemiol 5:343-356

el-Hazmi MA, Warsy AS, Addar MH (1992) DNA polymorphism in the beta-globin gene cluster in Saudi Arabs: relation to severity of sickle cell anaemia. Acta Haematologica 88:61-66

Fisher RA (1932) Statistical Methods for Research Workers, 4^{th} ed. Oliver & Boyd, London

Frankel WN and Schork NJ (1996) Who's afraid of epistasis. Nature Genetics 12:371-373

Fukuda K (1999) cdd+ release 0.76, Lausanne, Switzerland

Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, et al. (1983) A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306:234-238

Hollocher H, Templeton AR, DeSalle R, Johnston JS (1992) The molecular through eco-

logical genetics of abnormal abdomen. IV. Components of genetic-variation in a natural-population of Drosophila mercatorum. Genetics 130:355-366

Hollocher H and Templeton AR (1994) The molecular through ecological genetics of abnormal abdomen in Drosophila mercatorum. VI. The nonneutrality of the Y-chromosome rDNA polymorphism. Genetics 136:1373-1384

Huntington's Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72:971-983

Kempthorne, O (1957) Introduction to Genetic Statistics. Iowa State University Press, Aames Iowa

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Genet 22, 139-144

Lathrop GM, Ott J (1990) Analysis of complex diseases under oligogenic models and intrafamilial heterogeneity by the LINKAGE programs. Am J Hum Genet Suppl 47:A188

Motzkin TS, Raiffa H, Thompson GL, Thrall RM (1953) The double description method. In: Contributions to Theory of Games, Vol. 2, Kuhn HW and Tucher AW (eds.), Princeton University Press, Princeton, NJ, pp 51-73

Nelson MR, Kardia SLR, Ferrell RE, Sing CF (2001) A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res 11:458-470

Odenheimer DJ, Whitten CF, Rucknagel DL, Sarnaik SA, Sing CF (1983) Heterogeneity of sickle-cell anemia based on a profile of hematological variables. Am J Hum Genet 35:1224-1240

Risch N and Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516-1517

Ritchie M, Hahn L, Roodi N, Bailey L, Dupont W, Parl F, Moore J (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138-147

SAS Institute (1995) SAS release 6.11, Cary, NC

Schork NJ, Boehnke M, Terwilliger JD, Ott J (1993) Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. Am J Hum Genet 53:1127-1136

Sing CF and Davignon J (1985) Role of apolipoprotein E polymorphism in determining normal plasm lipid and lipoprotein variation. Am J Hum Genet 37:268-285

Spielman RS, McGinnis RE, and Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506-516

Templeton AR (2000) Epistasis and complex traits. In: Epistasis and the Evolutionary Process (eds. M. Wade, B. Brodie III, and J. Wolf). Oxford University Press, Oxford, pp 41-57

Woolf B (1955) On estimating the relation between blood group and disease. Ann Hum Genet 19:251-253

Wright S (1923) The roles of mutation, inbreeding, crossbreeding, and selection in evolution. Proc 6th Int Congress Genet 1:356-366

Zubenko GS, Hughes III HB, Stiffler JS (2001) D10S1423 identifies a susceptibility locus for Alzheimer's disease in a prospective, longitudinal, double-blind study of asymptomatic individuals. Molec Psych 6: 413-419

Table 1

Penetrances for a generic biallelic two-locus model

Genotypes	BB	Bb	bb	
AA	f_{11}	f_{12}	f_{13}	M_{A1}
Aa	f_{21}	f_{22}	f_{23}	M_{A2}
aa	f_{31}	f_{32}	f_{33}	M_{A3}
	M_{B1}	M_{B2}	M_{B3}	K

Table 2

Maximum heritability in purely epistasic two-locus models

K	p_A	p_B	V_I	h^2	K	p_A	p_B	V_I	h^2
0.50	0.5	0.5	0.250	1.000	0.40	0.5	0.5	0.170	0.708
		0.4	0.213	0.851			0.4	0.150	0.624
		0.3	0.241	0.964			0.3	0.158	0.655
		0.2	0.104	0.414			0.2	0.087	0.362
		0.1	0.052	0.207			0.1	0.020	0.083
	0.4	0.4	0.213	0.851		0.4	0.4	0.159	0.663
		0.3	0.212	0.849			0.3	0.153	0.639
		0.2	0.108	0.432			0.2	0.102	0.424
		0.1	0.052	0.207			0.1	0.033	0.136
	0.3	0.2	0.102	0.408		0.3	0.2	0.106	0.442
	0.2	0.2	0.057	0.226		0.2	0.2	0.057	0.236
		0.1	0.003	0.014			0.1	0.004	0.017
0.20	0.5	0.5	0.080	0.500	0.10	0.5	0.5	0.0200	0.222
		0.4	0.047	0.293			0.4	0.0103	0.114
		0.3	0.047	0.295			0.3	0.0118	0.131
		0.2	0.034	0.214			0.2	0.0086	0.095
		0.1	0.001	0.001			0.1	0.0003	0.004
	0.4	0.4	0.057	0.359		0.4	0.4	0.0148	0.164
		0.3	0.046	0.286			0.3	0.0104	0.115
		0.2	0.023	0.146			0.2	0.0051	0.057
		0.1	0.011	0.066			0.1	0.0005	0.006
	0.3	0.2	0.027	0.166		0.3	0.2	0.0043	0.048
	0.2	0.2	0.039	0.245		0.2	0.2	0.0022	0.024
		0.1	0.003	0.020			0.1	0.0023	0.026

Table 3

Two-locus penetrances yielding maximum h^2 for $K \in [0.25, 0.50]$

Genotypes	BB	Bb	bb
AA	$4K-1$	0	1
Aa	0	$2K$	0
aa	1	0	$4K-1$

Table 4

Two-locus penetrances yielding maximum h^2 for $K \in (0, 0.25]$

Genotypes	BB	Bb	bb
AA	0	0	$4K$
Aa	0	$2K$	0
aa	$4K$	0	0

Table 5

Three-locus penetrances yielding maximum h^2 for $K \in (0, 0.0625]$

	CC			Cc			cc		
	BB	Bb	bb	BB	Bb	bb	BB	Bb	bb
AA	0	0	$16K$	0	0	0	0	0	0
Aa	0	0	0	0	$4K$	0	0	0	0
aa	0	0	0	0	0	0	$16K$	0	0

Figure 1 Limits of two-locus, biallelic, purely epistatic ($V_A = V_D = 0$ at each locus) models with all alleles equally frequent. The bottom curve is the maximum variance due to genotype (V_I) for such models. The middle curve is the total variance as a function of disease prevalence ($V_T = K(1-K)$). The top curve is the maximum proportion of variance attributable to genotype ($h^2 = V_I/V_T$)) in such cases.

Figure 2 Limits of three-locus, biallelic, purely epistatic ($V_A = V_D = 0$ at each locus) models with all alleles equally frequent. The bottom curve is the approximate maximum variance due to genotype (V_I) for such models estimated using an iterative maximization algorithm from SAS. The middle curve is the total variance ($V_T = K(1-K)$) as a function of disease prevalence. The dots on the top curve are the maximum proportion of variance attributable to genotype ($h^2 = V_I/V_T$)) estimated by the iterative maximization method. The top curve represents the values for h^2 from particular models we have found.

Figure 3 Limits of three-locus, biallelic, purely epistatic ($V_A = V_D = 0$ at each locus) models with p_i = frequency of the less common allele at locus i. The bottom curve is the estimated maximum variance due to genotype (V_I) for such models. The middle curve is the total variance ($V_T = K(1-K)$) as a function of disease prevalence. The top curve is the estimated maximum proportion of variance attributable to genotype ($h^2 = V_I/V_T$)).

Figure 4 Limits of four-locus, biallelic, purely epistatic ($V_A = V_D = 0$ at each locus) models with all alleles equally frequent. The bottom curve is the estimated maximum variance due to genotype (V_I) for such models. The middle curve is the total variance ($V_T = K(1-K)$) as a function of disease prevalence. The top curve is the estimated maximum proportion of variance attributable to genotype ($h^2 = V_I/V_T$)).

Figure 5 Limits of four-locus, biallelic, purely epistatic ($V_A = V_D = 0$ at each locus) models with p_i = frequency of the less common allele at locus i. The bottom curve is the estimated maximum variance due to genotype (V_I) for such models. The middle curve is the total variance ($V_T = K(1-K)$) as a function of disease prevalence. The top curve is the estimated maximum proportion of variance attributable to genotype ($h^2 = V_I/V_T$)).

Figure 6 Comparison of three-locus biallelic, purely epistatic ($V_A = V_D = 0$ at each locus) models with and without two-locus interactions. All allele frequencies = 0.5. The top curve is the estimated maximum h^2 possible for purely epistatic three-locus models which permit two-locus marginal deviations. The middle curve provides a reference of the maximum h^2 possible for two-locus purely epistatic models. The bottom curve represents the estimated maximum h^2 possible in three-locus purely epistatic models with no two-locus marginal deviations.

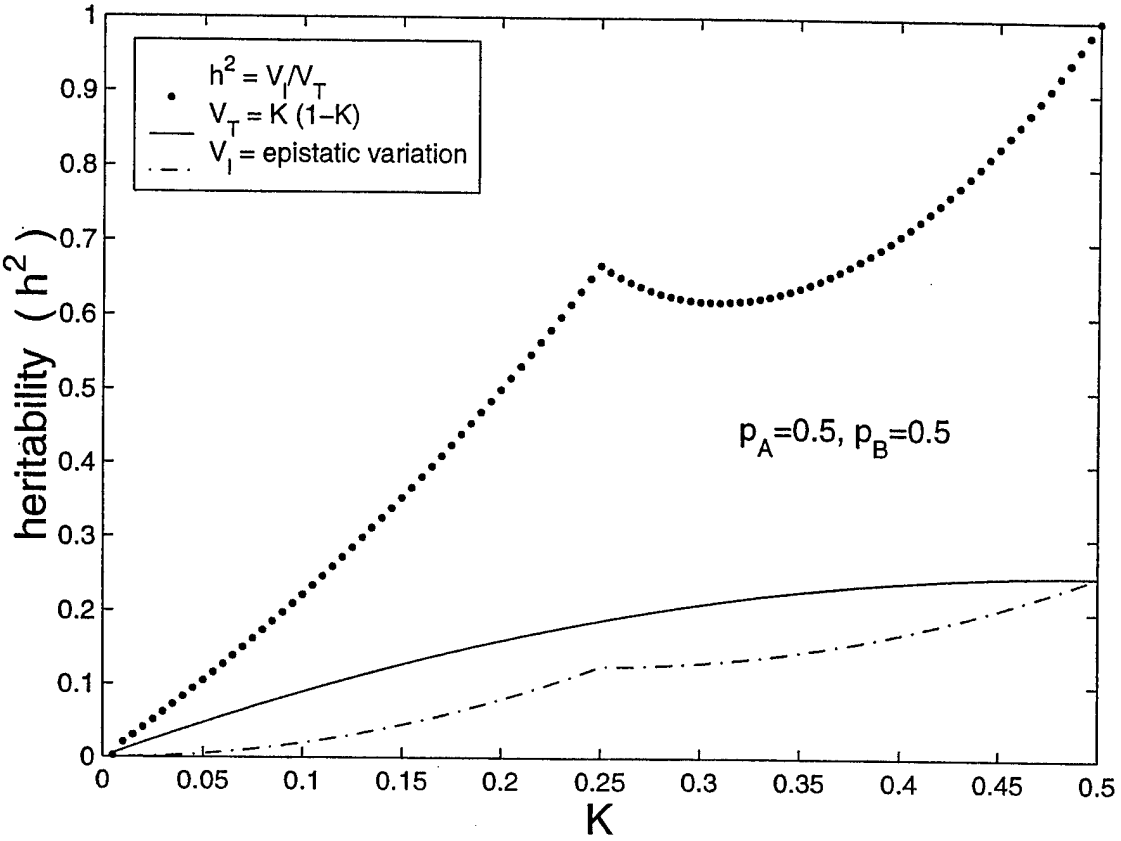
$h^2 = V_I/V_T$
$V_T = K(1-K)$
V_I = epistatic variation
p_A=0.5, p_B=0.5
heritability (h^2)
K
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.05
0.15
0.25
0.35
0.45

Figure 1

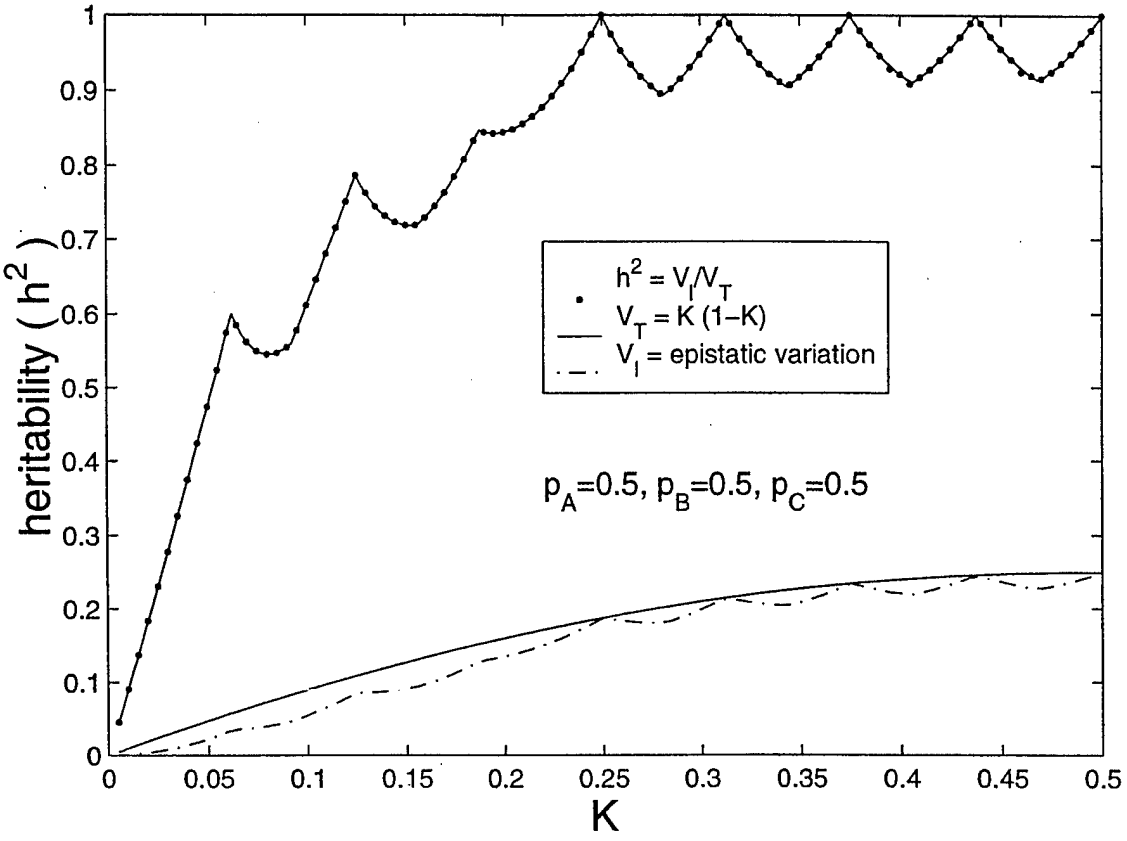
heritability (h^2)
K
$h^2 = V_I/V_T$
$V_T = K(1-K)$
V_I = epistatic variation
p_A=0.5, p_B=0.5, p_C=0.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.05
0.15
0.25
0.35
0.45

Figure 2

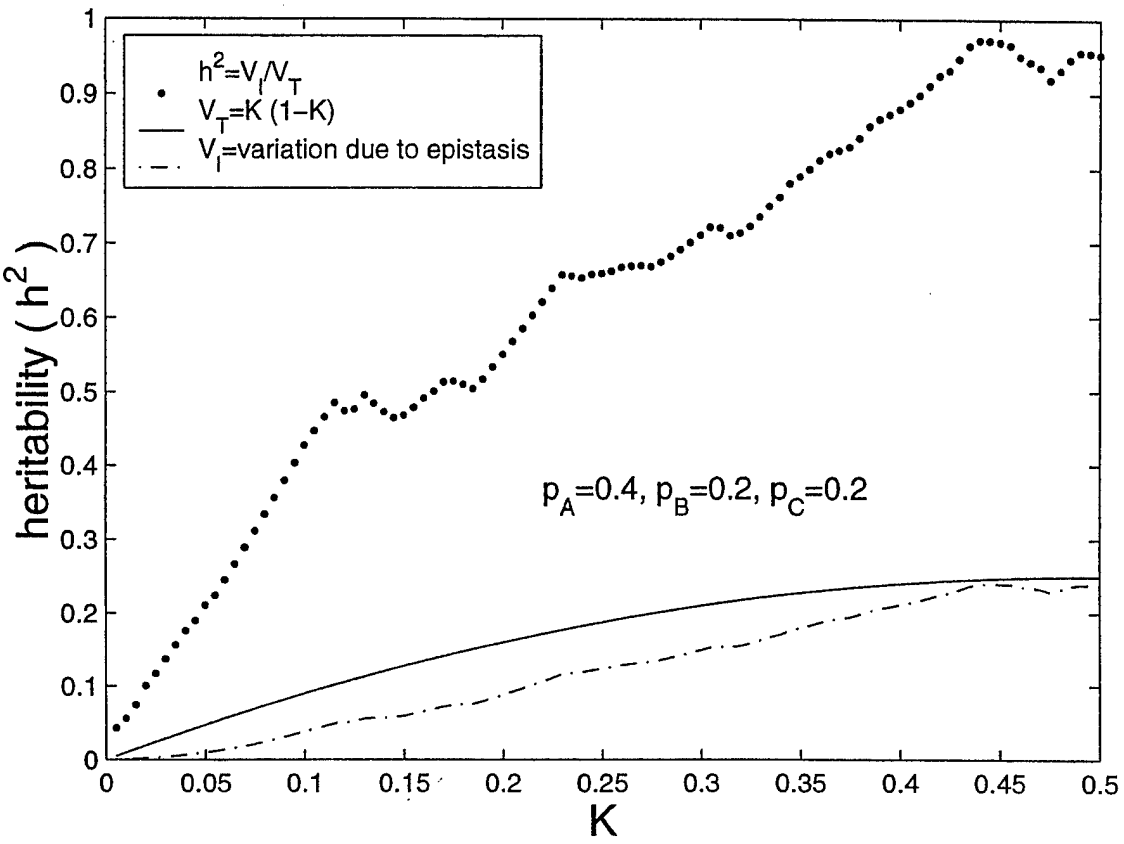
$h^2=V_I/V_T$
$V_T=K(1-K)$
V_I=variation due to epistasis
$p_A=0.4, p_B=0.2, p_C=0.2$
heritability (h^2)
K
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.05
0.15
0.25
0.35
0.45

Figure 3

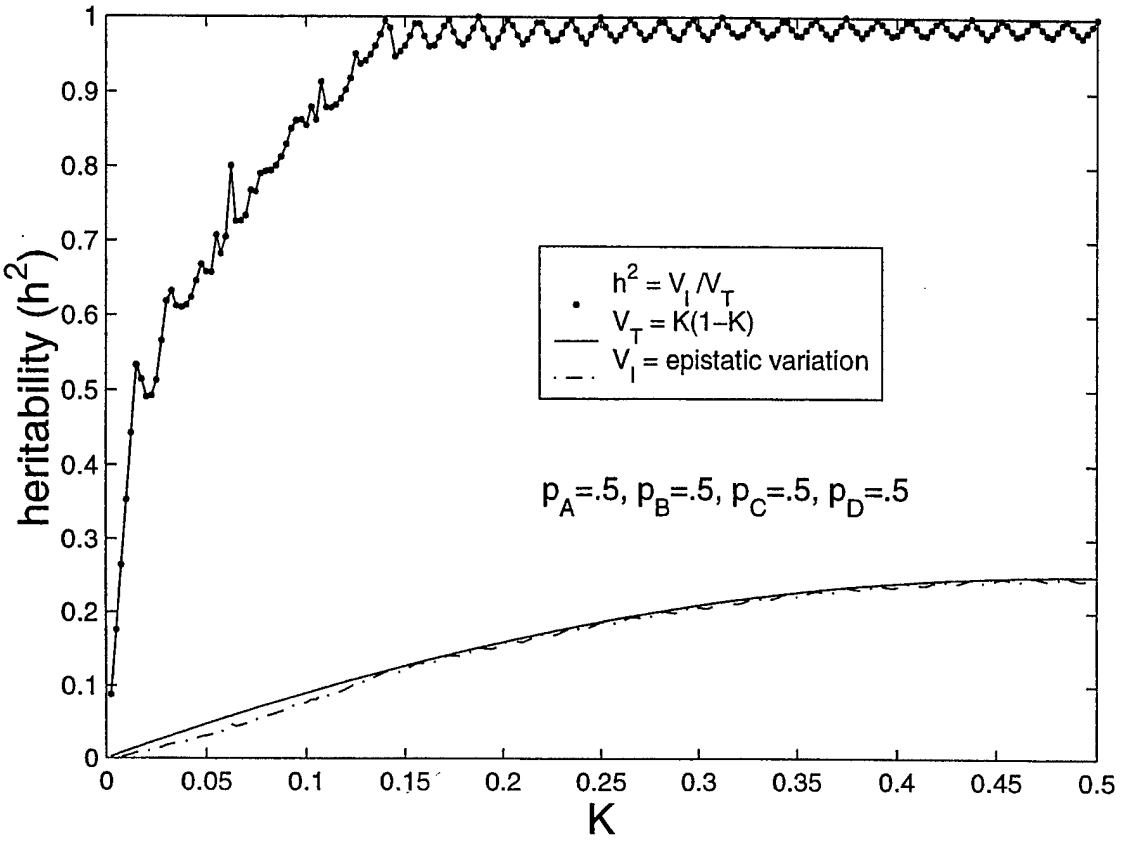
heritability (h^2)
K
$h^2 = V_I / V_T$
$V_T = K(1-K)$
V_I = epistatic variation
p_A=.5, p_B=.5, p_C=.5, p_D=.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.05
0.15
0.25
0.35
0.45

Figure 4

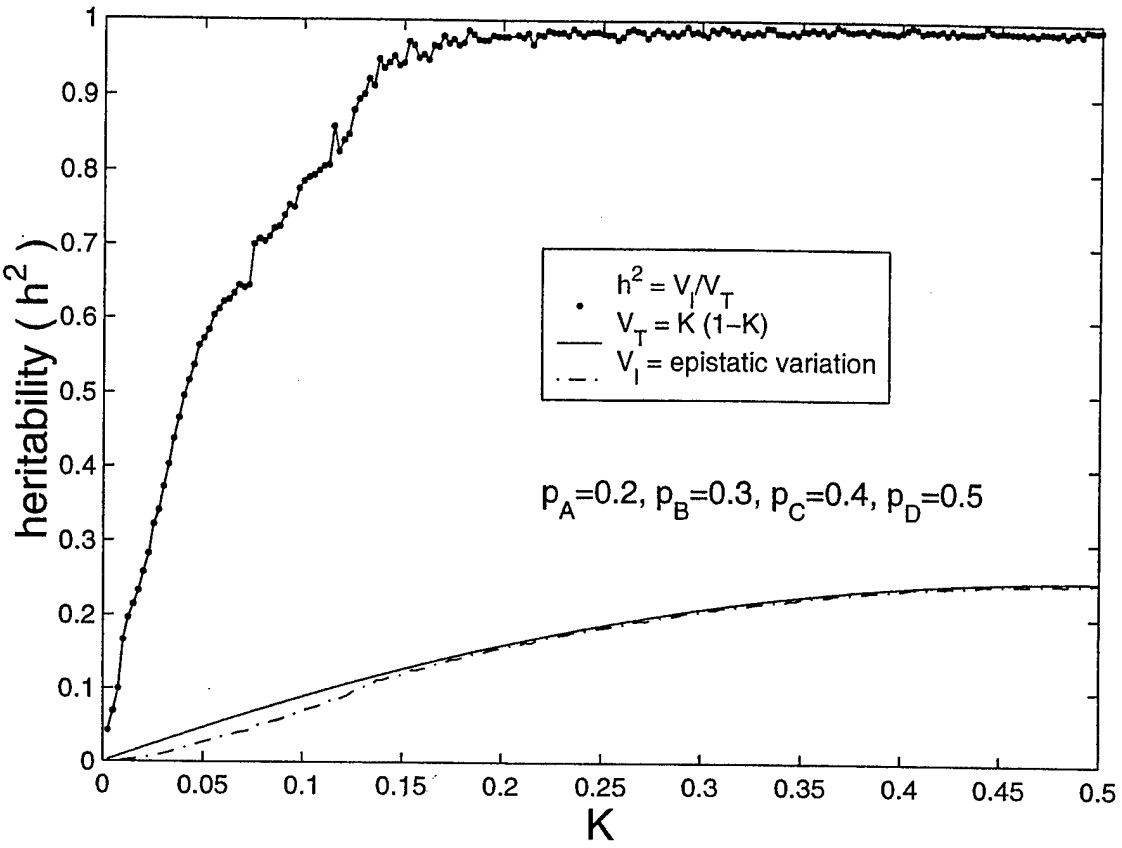
heritability (h^2)
K
$h^2 = V_I/V_T$
$V_T = K(1-K)$
V_I = epistatic variation
p_A=0.2, p_B=0.3, p_C=0.4, p_D=0.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.05
0.15
0.25
0.35
0.45

Figure 5

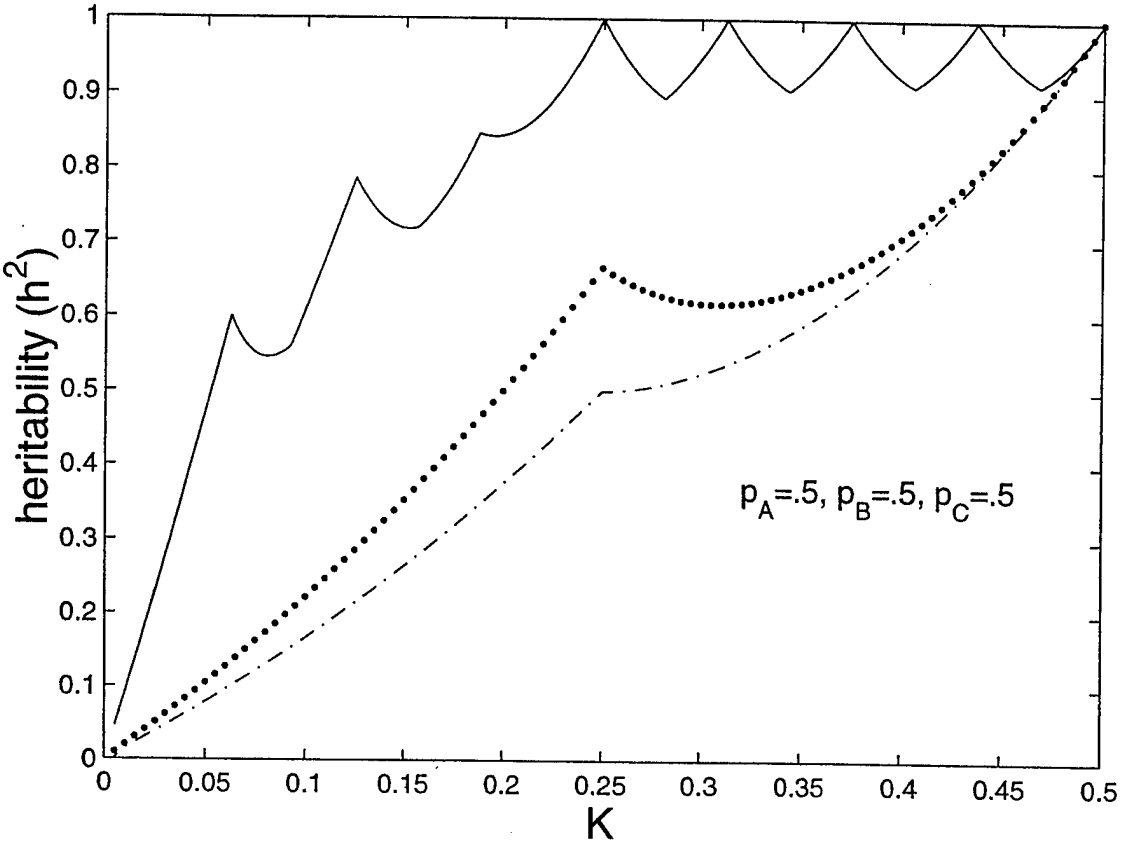
heritability (h^2)
K
p_A=.5, p_B=.5, p_C=.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.05
0.15
0.25
0.35
0.45

Figure 6

CAG and GGC Trinucleotide Repeats in the Androgen Receptor Gene and Prostate Cancer: The Long and the Short of it[1]

Brian K. Suarez [2], Jennifer Lin, William J. Catalona, Beth Haberer and Daniela S. Gerhard

Departments of Psychiatry [BKS, JL, DSG], Genetics [BKS,DSG], and Division of Urologic Surgery [BH,WJC], Washington University School of Medicine, St. Louis, MO, 63110.

Running title: *AR* polymorphism in prostate cancer

Key words: androgen receptor

prostate cancer

trinucleotide repeat

linkage disequilibrium

Gleason score

---

[1] Supported by awards from the Urological Research Foundation, the CaP CURE Foundation, USPHS grant MH31302 and DAMD17-00-1-0108 from the U. S. Army.

[2] To whom requests for reprints should be addressed, at Department of Psychiatry, Campus Box 8134, 660 S. Euclid, St. Louis, MO 63110. Phone: (314) 362-9433; Fax: (314) 747-1017; Email: bks@themfs.wustl.edu.

ABSTRACT

The androgen receptor (*AR*) is indispensable for maturation and maintenance of the prostate. It is encoded by a single copy X-linked gene. Two polymorphic trinucleotide repeats; $(CAG)_n$, resulting in a polyglutamine tract, and $(GGC)_n$, resulting in a polyglycine tract, are present in exon 1 and are widely believed to affect risk for the development of prostate cancer (CaP). To test this and other hypotheses, we genotyped these 2 polymorphisms in 287 multiplex CaP sibships and in 353 race-matched controls. We find no difference in the distribution of allele frequencies between cases and controls for either polymorphism and no evidence for linkage of the *AR* gene to prostate cancer. Allele size at either polymorphism is unrelated to age-at-diagnosis in our sample. There is, however, an increased probability of linkage disequilibrium between alleles at these two polymorphisms in CaP cases and evidence of an interaction that affects Gleason score in a small proportion of cases.

INTRODUCTION

The androgen receptor (*AR*)[3] is indispensable for the maturation and maintenance of the prostate. The *AR* is a member of the steroid receptor superfamily of ligand-dependent nuclear transcription factors that includes the receptors for glucocorticoids, estrogen, progesterone, vitamin D and mineralocorticoids. Proteins in this superfamily share a distinct structural similarity: an N-terminal modulatory domain, a highly conserved zinc-finger containing DNA binding domain, and a C-terminal hormone binding domain.

The human *AR* is encoded by a single copy gene located on Xq11-12 (1). Its open reading frame spans 8 exons (2,3), with exon 1 encoding the transactivating amino-terminal domain, exons 2 and 3 encoding the DNA binding domain, and exons 4-8 encoding the steroid binding domain. Three stretches of homopolymeric amino acids are encoded in exon 1, two of which are polymorphic: a polyglutamine $(CAG)_n$ tract that ends with an invariant CAA and a polyglycine tract, encoded by $(GGT)_3GGG(GGT)_2$ followed by a variable number of GGCs (4,5).

Three observations have bolstered the hypothesis that *AR* variation (particularly, variation in CAG repeat number) plays a key role in the pathophysiology of prostate cancer (CaP). (1) There is an inverse relationship between CAG repeat length and the *AR*'s transcriptional activity (6-10). (2) There is an inverse relationship between the prevalence of CaP among men of Asian, European and African ancestry and the mean number of CAG repeats found in these populations, respectively (11,12). (3) Somatic contractions in CAG repeat number in prostate cancer cells have been associated with tumor aggressiveness, cancer progression and failure of hormonal therapy

---

[3] The abbreviations used are: *AR*, androgen receptor; CaP, prostate cancer; PSA, prostate specific antigen; DRE, digital rectal examination; LD, linkage disequilibrium; GS, Gleason score; NPL, nonparametric linkage; SNP, single nucleotide polymorphism.

(13,14). Given these observations, and the fact that CaP develops only in the presence of androgens (males castrated prior to puberty do not develop CaP (15)), it is not surprising that a growing number of studies have sought to clarify the relationship between these polymorphisms and various aspects of CaP.

In this study, we report an analysis of the CAG and GGC polymorphisms in 287 multiplex CaP sibships and in 353 race-matched controls.

MATERIAL AND METHODS

Prostate Cancer Cases and Controls. Multiplex sibships were either ascertained from patients seen at Washington University School of Medicine by staff urologists, or were referred by other area urologists, or were participating in CaP support groups, or responded to our published solicitations. Two hundred and thirty of these families were included in our initial genome scan (16), 27 were added to our follow-up linkage study of chromosomes 1 and 16 (17), 22 were added for an analysis of polymorphisms in the CaP susceptibility gene, *HPC2/ELAC2* (18), and 8 families are new to this study.

The control subjects were ascertained from a large sample of men who have been followed for many years as part of a long-term prostate cancer screening study in which the subjects are screened at 6 to 12 month intervals with prostate specific antigen (PSA) blood tests and digital rectal examination (DRE) of the prostate (19). The large size of this pool allowed us to impose strict recruitment criteria. Briefly, to be enrolled as a control, the subjects were required to: 1) be at least 65 years old, 2) never have registered a PSA level in excess of 2.5 ng/ml (on a minimum of 3 occasions) nor had a DRE suspicious of CaP and, 3) have no known family history of CaP. This last criterion was operationalized by inquiring about the subject's brothers, father, grandfathers and maternal and paternal uncles. As a consequence of the first criterion, the control subjects are

significantly older than the CaP subjects (71.74 years vs. 65.33 years, $t=15.09$, $p<0.0001$). All subjects in this study are of European ancestry. The protocol of this study was approved by the Human Studies Committee of Washington University, and informed written consent was obtained from all participants.

Repeated Sampling of Multiplex Sibships. Unlike the standard case/control design, our cases consisted of sibships of affected brothers. We considered two options for testing hypotheses regarding the relationship between CAG and GGC repeat number and various parameters of CaP. First, we could draw a "one-time" sample by selecting one affected brother from each sibship. This option would yield a unique series of unrelated cases but the vagaries of sampling from the multiplex sibships could result in an unrepresentative sample. Alternatively, we could draw a sample, as above, but repeat the process many times to obtain an empirical distribution of the test statistic. We elected this second option and decided to repeat the sampling process 1,000 times for each hypothesis of interest.

Under the repeated sampling design, all affected cases in any particular replicate sample are independent from one another, but because approximately 8 to 19 percent of multiplex sibships contained only a single sib who was successfully genotyped for either (or both) markers, or had covariate data unavailable (Table 1), and because this single sib necessarily is always "randomly" selected, the 1,000 replicate samples have excess background nonindependence. In other words, there will be a minimum background overlap of about 8 to 19 percent between any two samples. To evaluate the influence, if any, contributed by these "singletons" families, two empirical distributions from 1,000 replicates were obtained for all hypotheses, first with all families included in the sampling frame and secondly with the singleton families excluded.

Statistical and Genetic Analysis. The Kolmogorov-Smirnov two-sample test implemented in the SAS (20) software package was used to evaluate the null hypothesis of no difference in the distribution of allele sizes in cases and controls. Maximum likelihood allele frequency estimates for the cases were obtained from the USERM13 subroutine of MENDEL (21,22). Because controls are unrelated to one another, their allele frequencies were obtained by direct gene-counting. To test various hypotheses regarding the association of allele sizes and CaP-related variables, we trichotomized both repeats into categories that can broadly be characterized as "short," "medium" and "long," to obtain as nearly as possible equal-sized categories based on the distribution of allele sizes in the control panel. The CAG repeat groupings (percentage in controls) were: ≤20 repeats (29.8%), 21-23 repeats (40.4%), and ≥24 repeats (29.8%). However, because the modal GGC repeat size contained more than half of the sample in controls (as well as cases), we were less successful in attaining an approximate tripartite uniform distribution for this polymorphism. For the GGC repeat the groupings were: <17 repeats (10.6%), 17 repeats (52.2%), and >17 repeats (37.2%). Two-way ANOVAs allowing for an interaction of the main effects as implemented in the GLM procedure of SAS (20) was used to evaluate hypotheses about age-at-diagnosis and Gleason score. Sample sizes for these tests are presented in Table 1.

[Table 1 about here]

Linkage Disequilibrium (LD) Analysis. Since the *AR* gene is located on the X-chromosome, determination of linkage phase in hemizygous males is unambiguously obtained by direct examination. Likelihood ratio chi-squares were used to assess the significance of the allelic associations in both cases and controls. In order to be included in the LD analysis, subjects had to be successfully genotyped for both repeats. Three hundred and forty-one control subjects and 596 CaP subjects met this criterion (Table 1).

***AR* Allele Sharing in Multiplex Families.** The evidence for linkage between CaP and the *AR* locus was evaluated with the computer program GENEHUNTER-PLUS (23,24).

CAG and GGC Repeat Length Assessment. The amplification of the 2 triplet polymorphisms in exon 1 of the *AR* gene needed special reagents because of the high GC content. The CAG polymorphism was optimally amplified with primers ADGRF (5'-tgcgcgaagtgatccagaac) and ADGRR (5'-cttggggagaaccatcctca) (25), though the second position of ADGRF primer has a "g" to conform to the canonical sequence (Genbank # XM_010429). The reaction utilized the GC-Melt kit (Clontech, CA) with amplification conditions specified by the manufacturer (94°C for 1 minute, then cycling between 94°C for 30 seconds and 68 °C for 3 minutes for 34 times with a final incubation at 68 °C for 3 minutes). The ADGRF primer was end-labeled by a standard T_4 kinase reaction. 37.5 ng of genomic DNA was amplified by 5 pmol of each primer in 1X reaction buffer (40 mM Tricine-KOH pH 9.2 at 25°C, 15 mM KOAc, 3.5 mM $Mg(OAc)_2$, 5% DMSO, 3.75 ug/mL BSA), 0.5M "GC-melt" proprietary reagent, 0.2 mM dNTPs and 1X Advantage-GC polymerase mix in a final volume of 8 µL. The same conditions were used for the GGC polymorphism using primers *AR*(GGC)F (acagccgaagaaggccagttgtat, end labeled) and *AR*(GGC)R (caggtgcggtgaagtcgctttcct) (26), though in some samples the alleles were difficult to call. Therefore, when FastStart (Roche, Germany) became available, we repeated 174 samples (15 of which were duplicates) using the FastStart DNA polymerase kit under conditions specified by the manufacturer (95 °C for 5 minutes, then 95 °C for 30 seconds, 65 °C for 30 seconds and 72 °C for 1 minute, the last 3 temperatures were repeated 34 times). 37.5 ng of genomic DNA was amplified with 5 pmol of each primer (the *AR*(GGC)F primer was end-labeled) in 1X reaction buffer (50 mM Tris-HCl pH 8.3 at 25°C, 10 mM KCl, 5 mM $(NH_4)_2SO_4$, 2 mM $MgCl_2$), 2X "GC-RICH solution"

proprietary reagent, 0.2 mM dNTPs and 0.4 u of FastStart *Taq* DNA polymerase in a final volume of 8 μL.

The PCR products were heated at 95°C for 15 minutes in 74% formamide and electrophoresed on 6% denaturing polyacrylamide gels for 2.5 hours, the gel dried and exposed to Amersham Hyperfilm MP X-ray film for 1-3 days. Two control DNAs and four M13 DNA sequencing lanes were included on each gel. The control DNAs included the CEPH samples genotyped by Marshfield Medical Research Foundation as well as the RPCI4-808O4 clone sequenced by the Sanger Centre (Genbank #AL049564). Two individuals, blind to familial relationships, called the allele sizes. The allele sizes were input independently into an Excel file and compared by a macro written for the purpose. There are 22 CAG repeats (66 bp) in the PAC clone resulting in a product of 207 bp. We also resequenced this region of the PAC and obtained the same number of CAG repeats. The RPCI4-808O4 clone has 17 GGCs. In total, at least 50% of all the genotypes were repeated at least twice.

RESULTS

Allele frequencies in cases and controls: Figures 1 and 2 display the cumulative distribution of allele frequencies in cases and controls for the CAG polymorphism and the GGC polymorphism, respectively. The null hypothesis that both cases and controls were sampled from the same population cannot be rejected (p=0.99 for both polymorphisms).

[Figures 1 and 2 about here]

Age-at-Diagnosis. The ANOVA analyses of age-at-diagnosis were unremarkable. Our cases provided no evidence of a main effect due to either the CAG or GGC repeat size category. Table 2 reports the results for the average difference (in percent) between the observed and predicted mean

age-at-diagnosis for 1,000 replicates sampled from all of the families and for 1,000 replicates sampled after removing the singleton families.

[Table 2 about here]

Gleason Score. The ANOVA analyses of Gleason scores yielded unexpected evidence for an interaction. When 1,000 random samples were drawn from all 271 families (singletons included), 48 gave evidence of a significant main effect at $p \leq 0.05$ (29 for CAG repeat size and 19 for GGC repeat size). This is close to the presumed type I error rate and is otherwise expected. A main effect is the average effect of the alleles at a gene taken over all other genes with which it may, or may not, interact. Two hundred and sixteen of the 1,000 replicate samples, however, showed a significant two-way interaction between the allele size categories of the two trinucleotide repeats. Fully 90 percent of these significant interactions occurred in replicate samples that did not yield a significant main effect for either CAG or GGC repeat size. An interpretation of this finding is that if a one-time random sample of unrelated cases were drawn from these 271 families, there is approximately a 21.6% chance of obtaining a significant CAG × GGC interaction.

When sampling was carried out after deleting the singleton families, 64 replicates showed a significant main effect—again close to the 5 percent expected by chance. One hundred and forty-seven replicates showed significant interactions (16 of which occurred in the presence of a significant main effect). Thus, 89 percent of these significant interactions occurred in the absence of a significant main effect.

Table 3 reports the percent deviation in Gleason score from that predicted (after accounting for the influence of the mostly non-significant main effects). The largest deviations occur in individuals with short GGC repeats. When all families are sampled, the Gleason score is 8.6

percent lower in individuals who also have a short CAG repeat and about 9-10% higher than predicted when the CAG allele falls in the medium or large category. Removing the 47 singleton families has no effect on the pattern although the size of the deviations is slightly greater.

[Table 3 about here]

Linkage disequilibrium. Both polymorphisms were successfully genotyped in 341 control subjects (Table 4). There is no evidence of significant linkage disequilibrium (X^2=7.099, p=0.131).

[Table 4 about here]

Both polymorphisms were successfully genotyped in 596 brothers with CaP (Table 1). Figure 3 reports two cumulative distributions of 1,000 X^2s computed on random replicate samples from our CaP families with singletons included (A) and singletons excluded (B).

[Figure 3 about here]

Both cumulative distributions provide evidence for increased linkage disequilibrium among CaP cases. Under the null hypothesis, we expect 50 of the 1,000 X^2 to be significant at approximately the α =5% level. The observed number of significant replicates (i.e., X^2>9.49, df=4) when all families are sampled (curve A) is 147, approximately 3 times more than expected. In other words, if a "one-time" random sample of unrelated cases were drawn from these multiplex families, there is approximately a 15 percent chance of finding significant linkage disequilibrium at the $\alpha \leq 0.05$ level. When the singleton families are excluded (curve B), 235 of the 1,000 replicate samples are significant at p<0.05—4.7 times the number expected under the hypothesis of no linkage disequilibrium. Accordingly, the excess number of significant replicates seen in curve A is not due to over-sampling the same 23 singletons.

Table 5 reports the average percent excess or decrement for each haplotype category in 1,000 replicate samples from the multiplex CaP families. When all families are sampled, there is

an approximate 25 percent increase in the frequency of short-GGC/short-CAG haplotypes, and a similar deficit in short-GGC/medium-CAG haplotypes. This pattern is accentuated when singleton families are deleted from the sampling.

[Table 5 about here]

Allele sharing at the *AR* locus in multiplex sibships. We observed no recombinants between the CAG and GGC repeats in our families. Accordingly, we treated the 68 haplotypes observed in these data as 68 alleles at a single locus and obtain a single-point NPL Z-score of -1.29, indicating a nonsignificant dearth of allele sharing.

DISCUSSION

Prostate cancer is the most common malignancy and the second leading cause of cancer-related death among American men. It is estimated that this year, 198,100 men will be newly diagnosed and about 31,500 will die of the disease (27). The disease has a complex and heterogeneous etiology involving multiple steps and multiple pathways to the malignant phenotype (28). Genetic changes that are sporadic events in somatic cells are not expected to increase the recurrence risk for prostate cancer among a proband's relatives. Nonetheless, virtually all family studies have found varying degrees of familiality for prostate cancer with the two strongest predictors of increased relative risk, aside from advanced age, being (1) the presence of multiple affected first degree relatives and (2) a proband with an unusually early age-of-onset (29).

Because of the androgen receptor's intrinsic role in the development of the prostate and maintenance of its physiological integrity in adulthood, it must be considered a strong candidate as a possible mediator of CaP susceptibility. If heritable variation in the *AR* gene mediates risk in some cases of CaP, then study of multiplex families should help clarify the *AR*'s role since multiplex families ought to contain proportionately fewer sporadic CaPs than a random sample of

incident cases. With this in mind, we genotyped affected brothers in 287 multiplex families and, to provide the greatest contrast, an exceptionally healthy race-matched control panel of 353 men.

A number of previous studies have reported a relationship between CAG repeat number and risk of CaP in European and European-derived samples, with shorter repeats conferring the greatest risk. Giovannucci et al. (30) found that men with shorter repeats were at particularly high risk for distant metastatic and fatal disease. Hakimi et al. (31) reported that short CAG repeats (<18) are over represented in men with lymph node-positive disease compared to either men with lymph node-negative disease or the general population. Since these workers found that the CAG allele frequency distribution in CaP cases was "remarkably comparable" to the distribution in the general population, they hypothesize that short CAG repeats may identify a subset of men at especially high risk. Stanford and colleagues (32) suggested a 3 percent decrease in CaP risk for each additional CAG repeat even though they observed "...no overall difference in the mean $(CAG)_n$ array length...between cases and controls." Ingles and associates (33) compared CAG repeat length in a small series of 57 CaP cases to a larger control series and concluded that those with fewer than 20 repeats experienced an approximately 2-fold increase in risk compared to men with 20+ repeats. A number of more recent studies, however, have failed to find any relationship between CAG repeat number and risk for CaP (34-37). In the new data reported here, we find no difference in the distribution of CAG allele sizes between CaP cases and controls and, in particular, no excess of short alleles in our multiplex families.

Fewer studies have assessed the relationship between the GGC repeat and CaP, perhaps because of the difficulty encountered in amplifying such a GC-rich region (38). Deletion of the polyglycine tract is known to reduce *AR* transcriptional activity by about 30 percent in transfection assays (39), but its role in CaP is poorly characterized. Hakimi et al., (31) reported that men with

fewer than 15 GGC repeats were overrepresented in their small sample of men undergoing radical retropubic prostatectomy, but unlike their finding for the polyglutamine tract, no association with lymph node status was observed. Stanford et al. (32) reported that men with fewer than 17 repeats have a significantly higher risk of CaP than men with longer repeats. Edwards et al. (36) found no relationship between GGC repeat number and overall risk to CaP. However, while Platz et al. (40) also found no difference in the mean number of GGC repeats in 582 CaP cases and 794 matched controls drawn from the U.S. Physicians' Health Study, they did report the presence of a quadratic relationship between the number of GGC repeats and risk to CaP. Risk was maximal at the mean and decreased 8% for every single repeat deviation in either direction from the mean. In the new data presented here, we find no difference in the distribution of allele sizes between cases and controls.

Regarding the possible association of either repeat with age-at-diagnosis, Hardy and colleagues (41) found a significant correlation between CAG repeat number and an early age-of-onset of CaP. Bratt et al. (34) also reported that shorter CAG repeats correlated with younger age-at-diagnosis, but only in men classified as having non-hereditary CaP. Neither Hakimi et al. (31) nor Stanford and colleagues (32) could find evidence of a relationship between CAG repeat number and age-at-onset. With respect to the GGC repeat, however, Stanford et al. (32) found increased risk in men aged 60-64 who had fewer than 17 repeats, compared to younger men with CaP. Our analysis of the relationship between the number of CAG and GGC repeats with age-at-diagnosis indicated no association for either polymorphism, singularly or jointly.

Our analysis of the relationship of the CAG and GGC repeats and Gleason score yielded an unexpected result: viz., the suggestion of an interaction, with lower than predicted Gleason scores in CaP cases whose repeats fall into the "short-short" category, and higher than predicted Gleason

scores for subjects whose GGC repeat is short but whose CAG repeat is medium or long. Clearly, caution needs to be exercised in interpreting the importance of this putative interaction. Thus, although 14.7% of the 1,000 replicates that excluded singletons and 21.6% of the 1,000 replicates that did not exclude singletons gave significant interactions, it must be born in mind that the replicates are not independent. Moreover, the significant interaction primarily involves a single category (the short GGC repeat class), which accounts for only about 10% of the cases. Under the hypothesis that men with short CAG alleles are at greater risk for the development of CaP and have a more aggressive disease—a hypothesis unconfirmed in our data—the suggestion that men with short alleles for both the CAG and GGC trinucleotide repeats have lower than predicted Gleason scores (GS) seems counterintuitive. To scrutinize the nature of this interaction, we divided the GSs into 4 functional categories ($\leq$4, 5-6, 7, and $\geq$8) that are believed to qualitatively capture differences in the aggressiveness of the disease (42) and used the CATMOD procedure of SAS (20) to predict the expected number of observations in each of the 4 GS categories using a main effects model. In the total data, we observed 20 cases whose haplotypes contain short alleles for both polymorphisms. The observed (and expected) numbers were GS $\leq$4 : 7, (2.6); GS= 5-6: 7, (8.6); GS=7: 6, (3.5); and GS$\geq$8: 0, (0.9). This analysis indicates that we observed more men in the short/short category than we would predict given the marginal frequencies (20 vs. 15.6)—consistent with the LD results—and that the greatest difference occurs for the lowest GS category. This analysis also reveals that the interaction, although significant in about 20 percent of the replicates, is based on small sample sizes and should not be over-interpreted. In this regard a recent report by Nam et al. (43) is of interest. While these authors were unable to demonstrate a global relationship between the number of CAG repeats (dichotomized at $\leq$18 versus >18 repeats) and biochemical disease recurrence (as assessed by PSA level) in 318 men who had undergone radical

prostatectomy for clinically localized CaP, they were able to demonstrate that short alleles were significantly overrepresented among a subgroup of men otherwise judged to be at low risk (i.e., Gleason score ≤ 6, stage pT2, and pretreatment PSA ≤ 10). Nam et al. (43) did not genotype their subjects for the GGC polymorphism. However, on the basis of the interaction suggested by our data, we would hypothesize that short GGC repeats are overrepresented in their "low risk" group (i.e., men who experience unexpected recurrence). Clearly, more research will be required to clarify whether the interaction suggested by our data can be replicated.

Linkage disequilibrium in a population refers to the nonrandom association of alleles at two or more loci. The ultimate source of LD is the occurrence of a new mutation, which immediately results in complete disequilibrium of the mutant allele with alleles at neighboring loci that perchance occupy the same chromosome. Over time, disequilibrium decays as a consequence of recombination. The approach to equilibrium for alleles on the non-pseudoautosomal segment of the X-chromosome is approximately twice as slow as for the autosomes, as recombination can only take place in females. Since the polyglutamine and polyglycine tracts are separated by only 1,110 bp (which is approximately equal to a recombination fraction of $\theta = 0.00001$), the search for linkage disequilibrium between these polymorphisms seems reasonable.

We found evidence of weak to moderate LD in replicate samples of unrelated CaP cases. The pattern of disequilibrium shown in Table 5 suggests an excess of "short-short" haplotypes among CaP cases. While the excess is relatively modest, it was present in 96 percent and 100 percent of all replicate contingency tables that yielded a significant X^2, for the entire sample and for the sample that excludes singletons, respectively. It is not clear, however, if this modest excess is a consequence of ascertaining multiplex CaP families, a subset of which are at particularly high risk due to an (unmeasured) genetic variant that is over represented on "short-short" haplotypes, or if it

is simply a property of the *AR* receptor in a European-derived population, since a nonsignificant excess in the same direction is also seen in the control sample. The regularity of the excess/deficit pattern for some of the other haplotypes is very pronounced and difficult to interpret. For instance, the largest deviation from expected numbers is the dearth of haplotypes containing short GGC repeats and medium CAG repeats, regardless of the inclusion/exclusion of the singleton families. A similar, albeit less pronounced, deficit is seen for haplotypes with short CAG repeats and medium GGC repeats. Again, it is worth noting that the same pattern is seen among the controls, so it is unlikely that either of these haplotypes confer protection against CaP.

Not all data sets show evidence of increased linkage disequilibrium for these two polymorphisms (31,32,36,44). Our linkage disequilibrium findings are similar to those reported by Irvine et al. (12) who observed no linkage disequilibrium in normal controls but moderate disequilibrium in CaP cases. Microsatellite alleles are known to mutate at rates far greater than individual nucleotides (45,46). Zhang et al. (47) amplified the polyglutamine polymorphism of individual human sperm to estimate the mutation rate of the CAG repeat. Among men with 20-22 CAG repeats, 9 mutations (3 expansions and 6 contractions) were seen in 685 sperm. The estimated mutation rate of 1.3×10^{-2} is approximately twice the estimate of 6.7×10^{-3} obtained from an analysis of CEPH families (26). Alleles in the high normal range (28-31 repeats), however, were 4.4 times more likely to undergo a mutation, again with contractions outnumbering expansions. For CAG repeats in the pathological range ($\geq$ 40 repeats)—sufficient to cause spinal and bulbar muscular atrophy in males—the mutation rate increases substantially (48). To our knowledge, no estimate of the germline mutation rate for the polyglycine tract has been made. Nonetheless, given the relatively high mutation rate for the CAG repeat, evidence for linkage disequilibrium between its alleles and alleles at the GGC repeat (despite their genomic propinquity)

in CaP cases seems *a priori* unlikely unless it is the result of ascertaining multiplex families whose increased risk is the result of an unmeasured polymorphism that is in weak LD with alleles at the two trinucleotide repeats. Interestingly, Ross et al. (49) report on a silent polymorphism at codon 211 in a sample of 208 African-American men. This SNP, which lies approximately half way between the two homopolyamino acid tracts, was found to be in disequilibrium with alleles at both the CAG and GGC repeats, although the latter were not in disequilibrium with one another.

A number of studies have estimated a higher risk for the development of CaP in brothers than fathers of probands (50,51). It is possible that the recent widespread use of PSA screening could have a larger impact on the apparent risk to brothers than fathers since, when these studies were conducted, many of the fathers were deceased and indolent CaP may have gone undetected. Both a secular trend for increasing CaP rates—perhaps due to greater longevity—as well as improved reporting, could also give rise to the appearance of an excess risk to brothers compared to fathers. From a purely genetic perspective, there are a number of transmission models compatible with such an observation. The presence of appreciable dominance variance, for instance, will result in higher risk to brothers than to fathers or, for that matter, a proband's sons (52). Since fathers and sons do not share an X chromosome, another mode of transmission compatible with increased risk to sibs is an X-linked susceptibility locus. And, indeed, at least one named (but as yet unidentified) X-linked susceptibility locus mapping to Xq27-28 (*HPCX*) has been inferred from linkage analysis (53). On the Marshfield female recombination map (http://www.marshfield.org/genetics) the peak signal reported by Xu et al., (53) is located approximately 90 cM from the *AR* locus. A number of linkage studies have genotyped markers in the vicinity of the *AR* locus, but none have yielded any significant evidence for excess allele sharing in multiplex families (37,54-57). In an earlier linkage analysis of 230 of the 287 multiplex

families included here (16), we obtained a negative multipoint NPL Z-score of -0.21 at DXS7132, the marker closest to the *AR* locus in our panel of 16 X-linked microsatellites. Using just the haplotypes determined by the CAG and GGC repeat polymorphisms, we obtain a single point NPL Z-score of -1.29 at the *AR* locus—indicating less allele sharing than expected by chance under the null hypothesis of no linkage. Our findings are similar to those reported by Sun et al. (54) who reported that only 18 of 41 brother pairs affected with CaP were concordant for the CAG repeat, and only 1 of 6 brother trios was concordant.

It is well known that broad "geographical races" show sizable variation in CaP rates (58-60). Rates are highest in populations with African ancestry, intermediate in European populations, and lowest in Asians. Edwards and colleagues (61) were the first to establish that these 3 geographical races also differ in the mean number of CAG repeats, although there is considerable overlap in the distributions. Coetzee and Ross (11) introduced the "CAG hypothesis" which posits that the difference in CaP rates observed between various geographical races could be explained, in part, by population differences in CAG repeat size. Since all of the subjects in our study are of European ancestry, we are unable to shed any light on this hypothesis. As reviewed above, however, the evidence from European and European-derived populations is mixed with respect to whether fewer CAG repeats are really associated with an increased risk for CaP. To our knowledge the only case-control study to investigate the "CAG hypothesis" in an Asian population was reported recently by Hsing et al. (38). They found that men with CAG repeats shorter than the median observed in the Shanghai region (i.e., 23 repeats) had a 65% increased risk of CaP. It is of interest that even in the positive reports from European and European-derived samples, the CAG effect is not seen except for repeat numbers much smaller than 23. We are unaware of any case-

control study of the relationship between CAG repeat size and CaP in an African or African-derived population.

The degree to which the population distribution of CAG repeats may have been shaped by stabilizing selection is also unknown. Even if men with short CAG repeats are at increased risk for the development of clinically significant disease—an hypothesis yet to be proven—since the vast majority will have completed reproduction by the time the disease develops, selection is unlikely to exert much of an effect on the low end of the CAG distribution. The recent suggestion that short CAG alleles protect females from the development of breast cancer (62) raises the intriguing possibility that selection exerts opposing effects on the Darwinian fitness of short CAG alleles in males and females.

There is some evidence that selection may be operating on the upper tail of the normal range of CAG repeat sizes. Two studies of subjects ascertained through infertility clinics indicate that men with CAG repeat lengths in the high normal range are at increased risk of infertility compared to controls. Tut et al. (63), for instance, report that 153 patients with defective sperm production were 4 times more likely to have longer CAG tracts ≥28 repeats) than controls (see also ref. 64). Dowsing et al. (65) found that men with idiopathic azoospermia or oligozoospermia had significantly longer CAG repeat tracts than controls. Two other studies, however, failed to find any association between CAG repeat length and infertility (66,67), although the sample sizes for both of these negative studies were exceedingly small. Any hypothesis that posits a role for stabilizing selection in the population distribution of CAG repeat sizes would also need to account for the racial variation in mean values.

We did not gather data regarding infertility from any of our CaP cases or from our controls. The self-report questionnaire administered to our CaP cases, however, did inquire about the number

of offspring they had fathered. Of the 273 ever-married CaP cases who answered this question, 16 reported having had no children. The mean CAG (22.06) and GGC (17.06) repeat numbers for this group did not differ from the CaP cases who reported one or more offspring (for CAG: $t=0.54$, $p=0.59$; for GGC: $t=0.07$, $p=0.94$). Moreover, among the CaP cases of proven fertility, there was no correlation between repeat length and the total number of offspring ($r=0.02$ for CAG, $p=0.73$, and $r = -0.06$ for GGC, $p=0.28$). Thus, while the absence of children is an imperfect indicator of fertility, our sample fails to provide any suggestion that men with larger repeats are less fertile.

In summary, we find no evidence for a difference in the distribution of allele sizes at either the CAG or the GGC repeat polymorphisms of the androgen receptor gene in men from multiplex CaP sibships and healthy controls. In our sample, there is no relationship between repeat number for either polymorphism and age-at-diagnosis. There is, however, evidence of an interaction between these polymorphisms and Gleason score. In randomized replication tests, CaP subjects whose poly-gly and poly-gln tracts fall in the "short-short" category were more likely to have Gleason scores lower than predicted while those in the short-medium and short-long categories were more likely to have Gleason scores higher than predicted. Additionally, randomized replication tests of CaP cases suggest an increased probability that alleles at these two polymorphisms are in disequilibrium with one another. In particular, a subset of short-GGC/short-CAG haplotypes may confer increased risk of CaP despite low Gleason scores. Finally, we find no evidence for linkage of the AR locus among brothers with CaP.

ACKNOWLEDGMENTS

We thank all of the subjects who have participated in this research. We gratefully acknowledge Kim Roehl, JoAnn Antenor and Sandy Schwartz for their help in subject recruitment and Scott Bloch, Niki Kesterson and Loan Nguyen for their assistance in the handling and genotyping of the DNA samples.

REFERENCES

1. Brown, C. J., Goss, S. J., Lubahn, D. B., Joseph, D. R., Wilson, E. M., French, F. S., and Willard, H. F. Androgen receptor locus on the human X chromosome: regional localization to Xq11-12 and description of a DNA polymorphism. Am. J. Hum. Genet., *44*:264-269, 1989.

2. Lubahn, D. B., Joseph, D. R., Sullivan, P. M., Willard, H. F., French, F. S., and Wilson, E. M. Cloning of human androgen receptor complementary DNA and localization to the X chromoosome. Science, *240*:327-330, 1988.

3. Lubahn, D.B., Brown, T.R., Simental, J.A., Higgs, H.N., Migeon, C.J., Wilson, E.M., and French, F.S. Sequence of the intron/exon junctions of the coding region of the human androgen receptor gene and identification of a point mutation in a family with complete androgen insensitivity. Proc. Natl. Acad. Sci., USA *86*:9534-9538, 1989.

4. Chang, C., Kokontis, J., and Liao, S.T. Structural analysis of cDNA and amino acid sequences of human and rat androgen receptors. Proc. Natl. Acad. Sci. USA, *85*:7211-7215, 1988.

5. Sleddens, H.F.B.M., Oostra, B.A., Brinkmann, A.O., and Trapman, J. Trinucleotide (GGN) repeat polymorphism in the human androgen receptor (*AR*) gene. Hum. Mol. Genet., *2*:493, 1993.

6. Mhatre, A.N., Trifiro, M.A., Kaufman, M., Kazemi-Esfarjani, P., Figlewicz, D., Rouleau, G., and Pinsky, L. Reduced transcriptional regulatory competence of the androgen receptor in X-linked spinal and bulbar muscular atrophy. Nat. Genet., *5*:184-188, 1993.

7. Chamberlain, N.L., Driver, E.D., and Miesfeld, R.L. The length and location of CAG trinucleotide repeats in the androgen receptor N-terminal domain affect transactivation function. Nucleic Acids Res., *22*:3181-3186, 1994.

8. Kazemi-Esfarjani, P., Trifiro, M.A., and Pinsky, L. Evidence for a repressive function of long polyglutamine tract in the human androgen receptor: possible pathogenetic relevance for the $(CAG)_n$-expanded neuronopathies. Hum. Mol. Genet., *4*:523-527, 1995.

9. Choong, C.S., Kemppainen, J.A., Zhou, Z.-X., and Wilson, E.M. Reduced androgen receptor gene expression with first exon CAG repeat expansion. Mol. Endocrin., *10*:1527-1535, 1996.

10. Beilin, J., Ball, E.M.A., Favaloro, J.M., and Zajac, J.D. Effect of the androgen receptor CAG repeat polymorphism on transcriptional activity: specificity in prostate and non-prostate cell lines. J. Mol. Endocrin., *25*:85-96, 2000.

11. Coetzee, G.A., and Ross, R.K. Re: Prostate cancer and the androgen receptor. J. Nat. Cancer Inst., *86*:872-873, 1994.

12. Irvine, R.A., Yu, M.C., Ross, R.K., and Coetzee, G.A. The CAG and GGC microsatellites of the androgen receptor gene are in linkage disequilibrium in men with prostate cancer. Cancer Res., *55*:1937-1940, 1995.

13. Schoenberg, M.P., Hakimi, J.M., Wang, S., Bova, G.S., Epstein, J.I., Fischbeck, K.H., Isaacs, W.B., Walsh, P.C., and Barrack, E.R. Microsatellite mutation (CAG24-->18) in the androgen receptor gene in human prostate cancer. Biochem. Biophys. Res. Commun., *198*:74-80, 1994.

14. Koivisto, P.A., and Rantala, I. Amplification of the androgen receptor gene is associated with P53 mutation in hormone-refractory prostate cancer. J. Pathol., *187*:237-241, 1999.

15. Wilding, G. The importance of steroid hormones in prostate cancer. Cancer Surv., *14*:113-130, 1992.

16. Suarez, B.K., Lin, J., Burmester, J.K., Broman, K.W., Weber, J.L., Banerjee, T.K., Goddard, A.B., Witte, J.S., Elston, R.C., and Catalona, W.J. A genome screen of multiples sibships with prostate cancer. Am. J. Hum. Genet., *66*:933-944, 2000.

17. Suarez, B.K., Lin, J., Witte, J.S., Conti, D.V., Resnick, M.I., Klein, E.A.,Burmester, J.K., Vaske, D.A., Banerjee, T.K., and Catalona, W.J. Replication linkage study for prostate cancer susceptibility genes. Prostate, *45*:106-114, 2000.

18. Suarez, B.K., Gerhard, D.S., Lin, J., Haberer, B., Nguyen, L., Kesterson, N.K., and Catalona, W.J. Polymorphisms in the prostate cancer suceptibility gene HPC2/ELAC2 in multiplex families and healthy controls. Cancer Res., *61*:4982-4984, 2001.

19. Smith, D.S., Humphrey, P.A., and Catalona, W.J. The early detection of prostate carcinoma with prostate specific antigen: the Washington University experience. Cancer (Phila.), *80*:1852-1856, 1997.

20. SAS Institute Inc. The SAS System, Version 6.09. Cary, NC: SAS Institute Inc., 1992.

21. Lange, K., Weeks, D., and Boehnke, M. Programs for pedigree analysis: MENDEL, FISHER, and dGENE. Genet. Epidemiol., *5*:471-472, 1988.

22. Boehnke, M. Allele frequency estimation from data on relatives. Am. J. Hum. Genet., *48*:22-25, 1991.

23. Kruglyak, L., Daly, M.J., Reeve-Daly, M.P., and Lander, E.S. Parametric and nonparametric linkage analysis: a unified multipoing approach. Am. J. Hum. Gene., *58*:1347-1363, 1996.

24. Kong, A., and Cox, N.J. Allele-sharing models: LOD scores and accurate linkage tests. Am. J. Hum. Genet., *61*:1179-1188, 1997.

25. Sleddens, H.F., Oostra, B.A., Brinkmann, A.O., Trapman, J. Trinucleotide repeat polymorphism in the androgen receptor gene (*AR*). Nucleic Acids Res., *20*:1427, 1992.

26. Donnelly, A., Koxman, H., Gedeon, A. K., Webb, S., Lynch, M., Sutherland, G. R., Richards, R. I., and Mulley, J. C. A linkage map of microsatellite markers of the human X chromosome. Genomics, *20*:363-370, 1994.

27. Greenlee, R.T., Hill-Harmon, M. B., Murray, T., and Thun, M.. Cancer statistics, 2001. Cancer J. Clin., *51*: 15-36, 2001.

28. Isaacs, W.B., and Bova, G.S. Prostate cancer. In: Vogelstein, B., Kinzler, K. [eds.] The Genetic Basis of Human Cancer, pp 653-660. New York, McGraw-Hill, 1998.

29. Keetch, D.W., Rice, J.P., Suarez, B.K., and Catalona, W.J. Familial aspects of prostate cancer: A case control study. J. Urol., *154*:2100-2102, 1995.

30. Giovannucci, E., Stampfer, M.J., Krithivas, K., Brown, M., Brufsky, A., Talcott, J., Hennekens, C.H., and Kantoff, P.W. The CAG repeat within the androgen receptor gene and its relationship to prostate cancer. Proc. Natl. Acad. Sci. USA, *94*: 3320-3323, 1997.

31. Hakimi, J.M., Schoenberg, M.P., Rondinelli, R.H., Piantadosi, S., and Barrack, E.R. Androgen receptor variants with short glutamine or glycine repeats may identify unique subpopulations of men with prostate cancer. Clin. Cancer. Res., *3*:1599-1608, 1997.

32. Stanford, J.L., Just, J.J., Gibbs, M., Wicklund, K.G., Neal, C.L., Blumenstein, B.A., and Ostrander, E.A. Polymorphic repeats in the androgen receptor gene: molecular markers of prostate cancer risk. Cancer Res., 57:1194-1198, 1997.

33. Ingles, S.A., Ross, R.K., Yu, M.C., Irvine, R.A., La Pera, G., Haile, R.W.,and Coetzee, G.A. Association of prostate cancer risk with genetic polymorphisms in vitamin D receptor and androgen receptor. J. Natl. Cancer Inst., *89*:166-170, 1997.

34. Bratt, O., Borg, A., Kristoffersson, U., Lundgren, R., Zhang, Q.-X., and Olsson, H. CAG repeat length in the androgen receptor gene is related to age at diagnosis of prostate cancer

and response to endocrine therapy, but not to prostate cancer risk. Br. J. Cancer, *81*:672-676, 1999.

35. Correa-Cerro, L., Wohr, G., Haussler, J., Berthon, P., Drelon, E., Mangin, P., Fournier, G., Cussenot, O., Kraus, P., Just, W., et al. $(CGA)_nCAA$ and GGN repeats in the human androgen receptor gene are not associated with prostate cancer in French-German population. Euro. J. Hum. Genet., *7*: 357-362, 1999.

36. Edwards, S.M., Badzioch, M.D., Minter, R., Hamoudi, R., Collins, N., Ardern-Jones, A., Dowe, A., Osborne, S., Kelly, J., Shearer, R., Easton, D.F., Saunders, G.F., Dearnaley, D.P., and Eeles, R.A. Androgen receptor polymorphisms: Association with prostate cancer risk, relapse and overall survival. Int. J. Cancer (Pred. Oncol.), *84*:458-465, 1999.

37. Lange, E.M., Chen, H., Brierley, K., Livermore, H., Wojno, K.J., Langefeld, C.D., Lange, K., and Cooney, K.A. The polymorphic exon 1 androgen receptor CAG repeat in men with a potential inherited predispostion to prostate cancer. Cancer Epidemiol. Biomark. Prev., *9*:439-442, 2000.

38. Hsing, A.W., Gao, Y.-T., Wu, G., Wang, X., Deng, J., Chen, Y.-L., Sesterhenn, I.A., Mostofi, F.K., Benichou, J., and Chang, C. Polymorphic CAG and GGN repeat lengths in the androgen receptor gene and prostate cancer risk: A population-based case-control study in China. Cancer Res., *60*:5111-5116, 2000.

39. Gao, T., Marcelli, M., and McPhaul, M.J. Transcriptional activation and transient expression of the human androgen receptor. J. Steroid Biochem. Mol. Biol., *59*:9-20, 1996.

40. Platz, E.A., Giovannucci, E., Dahl, D.M., Krithivas, K., Hennekens, C.H., Brown, M., Stampfer, M.J., and Kantoff, P.W. The androgen receptor gene GGN microsatellite and prostate cancer risk. Cancer Epidemiol. Biomark. Prev., *7*:379-384, 1998.

41. Hardy, D.O., Scher, H.I., Bogenreider, T., Sabbatini, P., Zhang, Z.-F., Nanus, D.M., and Catterall, J.F. Androgen receptor CAG repeat lengths in prostate cancer: correlation with age of onset. J. Clin. Endocrinol. Metab., *81*:4400-4405, 1996.

42. Epstein, J. I., Pizov, G., and Walsh, P. C. Correlation of pathological findings with progression after radical retropubic prostatectomy. Cancer, *71*: 3582-3593, 1993.

43. Nam, R.K., Elhaji, Y., Krahn, M.D., Hakimi, J., Ho, M., Chu, W., Sweet, J., Trachtenberg, J., Jewett, M.A., and Narod, S.A. Significance of the CAG repeat polymorphism of the androgen receptor gene in prostate cancer progression. J. Urol., *164*:567-572, 2000.

44. Macke, J.P., Hu, N., Hu, S., Bailey, M., King, V.L., Brown, T., Hamer, D., and Nathans, J. Sequence variation in the androgen receptor gene is not a common determinant of male sexual orientation. Am. J. Hum. Genet., *53*:844-852, 1993.

45. Weber, J. L. and Wong, C. Mutation of human short tandem repeats. Hum. Molec. Genet., 2:1123-1128, 1993.

46. Suarez, B. K., Hampe, C. L., Zambuto, C., and Matise, T. C. Microsatellite motif and allele frequency divergence. Am. J. Hum. Genet., *61*:A44, 1997.

47. Zhang, L., Leeflang, E.P., Yu, J., and Arnheim, N. Studying human mutations by sperm typing: instability of CAG trinucleotide repeats in the human androgen receptor gene. Nat. Gen., 7:531-535, 1994.

48. La Spada, A.R., Roling, D.B., Harding, A.E., Warner, C.L., Spiegel, R., Hausmanowa-Petrusewicz, I., Yee, W.-C., and Fischbeck, K.H. Meiotic stability and genotype-phenotype correlation of the trinucleotide repeat in X-linked spinal and bulbar muscular atrophy. Nat. Genet., 2:301-304, 1992.

49. Ross, R.K., Pike, M.C., Coetzee, G.A., Reichardt, J.K.V., Yu, M.C., Feigelson, H., Stanczyk, F.Z., Kolonel, L.N., and Henderson, B.E. Androgen metabolism and prostate cancer: establishing a model of genetic susceptibility. Cancer Res., *58*:4497-4504, 1998.

50. Narod, S.A., Dupont, A., Cusan, L., Diamond, P., Gomez, J.L., Suburu, R., and Labrie, F. The impact of family history on early detection of prostate cancer. Nat. Med., *1*:99-101, 1995.

51. Monroe, K.R., Yu, M.C., Kolonel, L.N., Coetzee, G.A., Wilkens, L.R., Ross, R.K., and Henderson, R.E. Evidence of an X-linked or recessive genetic component to prostate cancer risk. Nat. Med., *1*:827-829, 1995.

52. Suarez, B. K., Reich, T., and Trost, J. Limits of the general two-allele locus model with incomplete genetrance. Ann. Hum. Genet., *40*:231-244, 1976.

53. Xu, J., Myers, D., Freije, D., Isaacs, S., Wiley, K., Nusskern, D., Ewing, C., Wilkens, E., Bujnovszky, P., Bova, G. S, et al. Evidence for a prostate cancer susceptibility locus on the X chromosome. Nat. Genet., *20*:175-179, 1998.

54. Sun, S., Narod, S.A., Aprikian, A., Ghadirian, P., and Labrie, F. Androgen receptor and familial prostate cancer. Nat. Med., *1*:848-849, 1995.

55. Smith, J. R., Freije, D., Carpten, J. D., Gronberg, H., Xu, J., Isaacs, S. D., Brownstein, M. J., Bova, G. S., Guo, H., Bujnovszky, P. et al. Major susceptibility locus for prostate cancer on chromosome 1 suggested by a genome-wide search. Science, *274*:1371-1374, 1996.

56. Gibbs, M., Stanford, J. L., Jarvik, G. P., Janer, M., Badzioch, M., Peters, M. A., Goode, E. L., Kolb, S., Chakrabarti, L., Shook, M., et al. A genomic scan of families with prostate cancer identifies multiple regions of interest. Am. J. Hum. Genet., *67*:100-109, 2000.

57. Hsieh, C-L., Oakley-Girvan, I., Balise, R. R., Halpern, J., Gallagher, R. P., Wu, A. H., Kolonel, L. N., O'Brien, L. E., Lin, I. G., Van Den Berg, D. J., et al. A genome screen of families with multiple cases of prostate cancer: Evidence of genetic heterogeneity. Am. J. Hum. Genet., *69*:148-158, 2001.

58. Parkin, D.M., Pisani, P., and Ferlay, J. Estimates of the world-wide incidence of 18 major cancers in 1985. Int. J. Cancer, *54*:594-606, 1993.

59. Whittemore, A.S., Wu, A.H., Kolonel, L.N., John, E.M., Gallagher, R.P., Howe, G.R., West, D.W., Teh, C.Z., and Stamey, T. Family history and prostate cancer risk in black, white, and Asian men in the United States and Canada. Am. J. Epidemiol., *141*:732-740, 1995.

60. Platz, E.A., Rimm, E.B., Willett, W.C., Kantoff, P.W., and Giovannucci, E. Racial variation in prostate cancer incidence and in hormonal system markers among male health professionals. J. Nat. Cancer Inst., *92*:2009-2017, 2000.

61. Edwards, A., Hammond, H.A., Jin, L., Caskey, C.T., and Chakraborty, R. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics, *12*:241-253, 1992.

62. Giguère, Y., Dewailly, E., Brisson, J., Aystte, P., Laf lamme, N., Demers, A., Forest, V.-I., Dodin, S., Robert, J., and Rousseau, F. Short polyglutamine tracts in the androgen receptor are protective against breast cancer in the general populations. Cancer Res., *61*: 5869-5874, 2001.

63. Tut, T.G., Ghadessy, F.J., Trifiro, M.A., Pinsky, L., and Yong, E.L. Long polyglutamine tracts in the androgen receptor are associated with reduced trans-activation, impaired sperm production, and male infertility. J. Clin. Endocrin. Metab., *82*:3777-3782, 1997.

64. Yong, E. L., Ghadessy, F., Wang, Q., Mifsud, A., and Ng, S. C. Androgen receptor transactivation domain and control of spermatogenesis. Rev. Reprod., *3*:141-144, 1998.

65. Dowsing, A.T., Yong, E.L., Clark, M., McLachlan, R.I., de Kretser, D.M., and Trounson, A.O. Linkage between male infertility and trinucleotide repeat expansion in the androgen-receptor gene. Lancet, *354*:640-643, 1999.

66. Puscheck, E. E., Behzadian, M. A., and McDonough, P. G. The first analysis of exon 1 (the transactivation domain) of the androgen receptor gene in infertile men with oligospermia or azoospermia. Fertil. Steril., *62*:1035-1038, 1994.

67. Tincello, D.G., Saunders, P. T., and Hargreave, T. B. Preliminary investigations on androgen receptor gene mutations in infertile men. Mol. Hum. Reprod., *3*:941-943, 1997.

Table 1. Sample size for various partitions of the family data set. All sibships are multiplex (ie, contain at least 2 affected brothers), but some variables are missing for some subjects. Thus, singletons have at least one (unmeasured) brother with CaP.

Sibship Size	CAG	GGC	CAG & GGC	Age-at-Diagnosis	Gleason Score
Singleton	20	16	23	51	47
Pairs	215	218	213	179	183
Trios	46	49	45	39	39
Quartet	4	3	3	2	2
Total Families	285	286	284	271	271
Total Individuals	604	611	596	534	538

Table 2. Mean percent excess or decrement in age-at-diagnosis as a function of CAG and GGC repeat size in the *AR* gene based on 1,000 random samples of multiplex sibships. See text for definition of "short" "medium" and "long."

	All Families (N=271)			Singletons Removed (N=220)		
CAG Repeat Size	GGC Repeat Size			GGC Repeat Size		
	Short	Medium	Long	Short	Medium	Long
Short	-0.1	-0.4	0.0	-0.3	-0.2	0.2
Medium	-1.9	0.4	0.2	-3.3	1.2	-0.8
Long	4.0	-0.5	-0.8	4.4	-0.8	-0.7

Table 3. Mean percent excess or decrement in Gleason score as a function of CAG and GGC repeat size in the *AR* gene based on 1,000 random samples of multiplex sibships. See text for definition of "short" "medium" and "long."

	All Families (N=271)			Singletons Removed (N=224)		
CAG Repeat Size	GGC Repeat Size			GGC Repeat Size		
	Short	Medium	Long	Short	Medium	Long
Short	-8.6	-1.1	2.7	-10.1	-1.3	2.8
Medium	9.8	-0.3	0.4	13.5	0.5	0.3
Long	9.1	1.4	-6.4	11.1	0.0	-4.4

Table 4. Joint distribution of the CAG and GGC repeat sizes in the first exon on the *AR* gene in 341 unrelated controls. See text for definition of "short" "medium" and "long."

CAG Repeat Size	GGC Repeat Size		
	Short	Medium	Long
Short	15	43	45
Medium	12	77	49
Long	9	58	33

Table 5. Mean percent excess or decrement for nine categories of haplotypes from 1,000 random samples of multiplex CaP sibships. See text for definition of "short" "medium" and "long."

	All Families (N=284)			Singletons Removed (N=261)		
CAG Repeat Size	GGC Repeat Size			GGC Repeat Size		
	Short	Medium	Long	Short	Medium	Long
Short	26.7	-17.8	18.3	44.0	-14.0	10.4
Medium	-28.4	7.2	-3.8	-56.6	6.9	1.5
Long	7.4	10.7	-15.8	22.5	7.5	-14.3

Figure 1. Cumulative distribution of CAG allele frequencies for 604 CaP cases and 352 controls.

Figure 2. Cumulative distribution of GGC allele frequencies for 611 CaP cases and 341 controls.

Figure 3. Cumulative distribution of likelihood ratio chi-squares for 1000 replicate samples from all 284 sibships (A) and from 261 sibships (B) with 2 or more genotyped sibs. Each chi-square is a test of the hypothesis of no association between CAG and GGC alleles when each polymorphism is trichotomized (see text for size ranges). The arrow marks the value of the chi-square obtained from 341 control subjects. The vertical dashed line marks the position of the α =5% critical value on 4 degrees of freedom. 14.7% and 23.5% of the replicate chi-squares from A and B, respectively, exceed the 5% critical value.

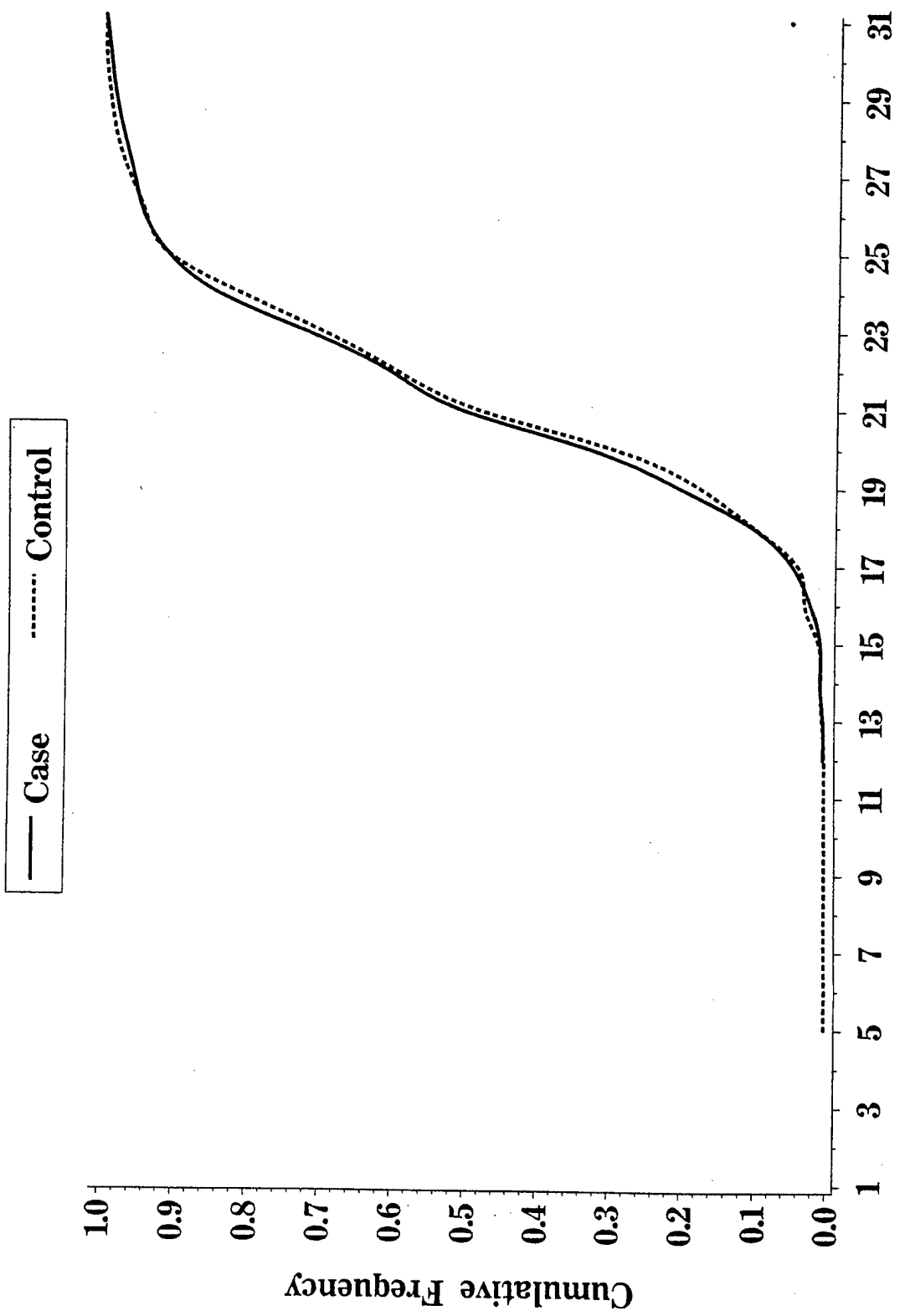
Case
Control
Cumulative Frequency
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
CAG Repeats

FIGURE 1

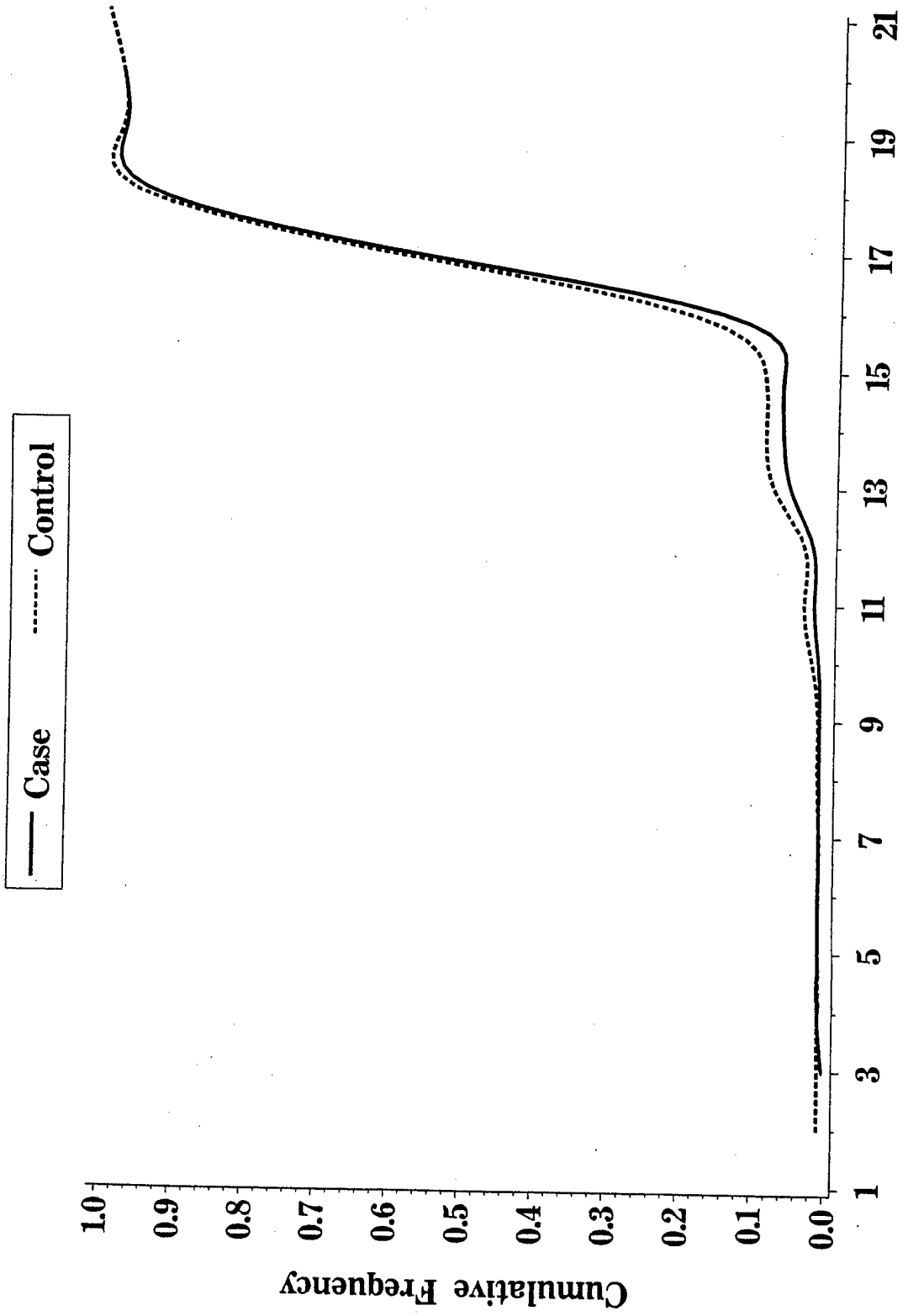
Case
Control
Cumulative Frequency
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
3
5
7
9
11
13
15
17
19
21
GGC Repeats

FIGURE 2

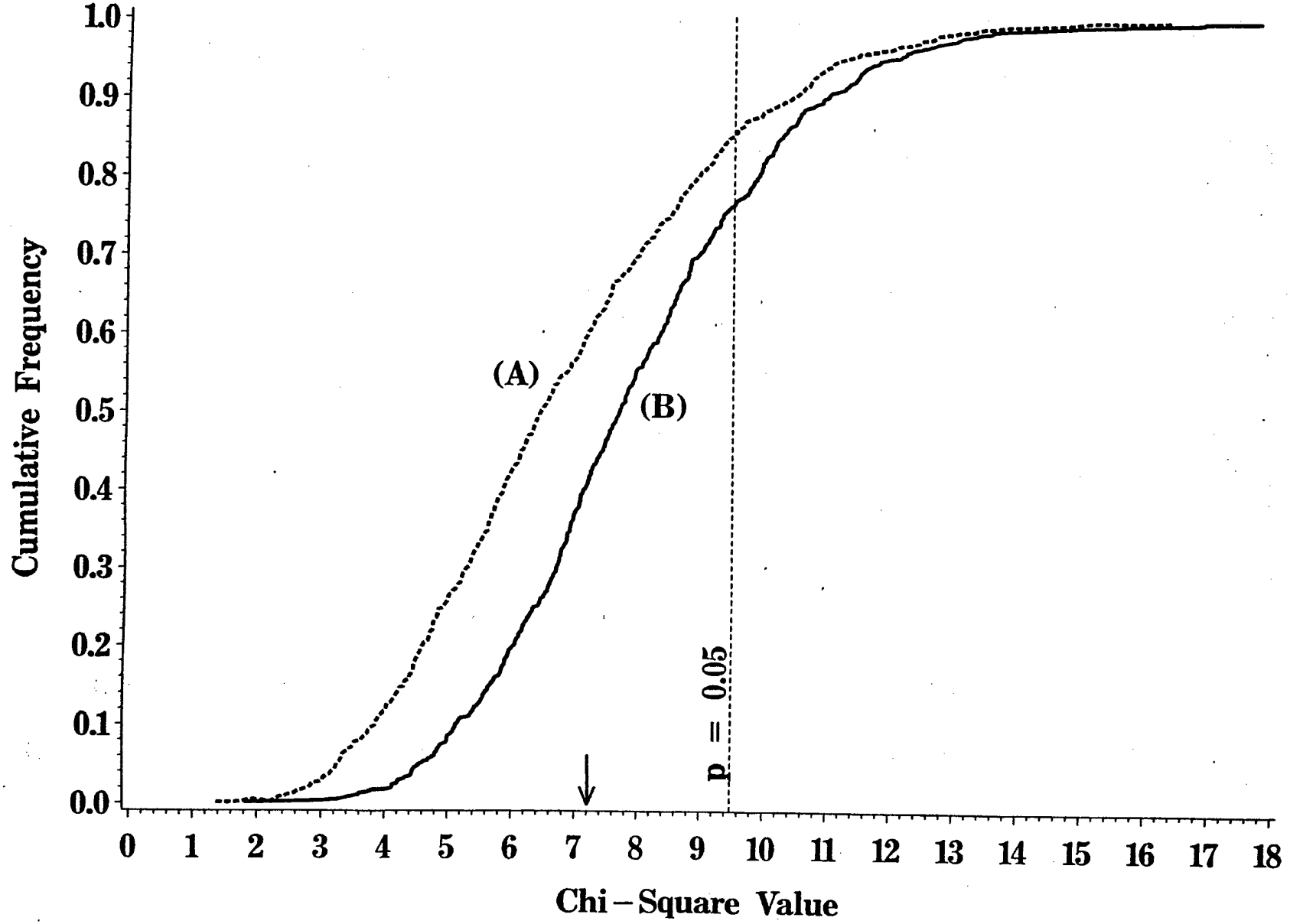
Cumulative Frequency
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
(A)
(B)
p = 0.05
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Chi – Square Value

FIGURE 3